

For Reference

NOT TO BE TAKEN FROM THIS ROOM

For Reference

NOT TO BE TAKEN FROM THIS ROOM

Ex LIBRIS
UNIVERSITATIS
ALBERTAENSIS



Regulations Regarding Theses and Dissertations

[illegible]



Digitized by the Internet Archive
in 2018 with funding from
University of Alberta Libraries

<https://archive.org/details/Jackson1964>

Thesis
1964
#55

THE UNIVERSITY OF ALBERTA

RATIONAL APPROXIMATIONS

by

Garry R. Jackson

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICS

EDMONTON, ALBERTA

DECEMBER, 1963

1964

UNIVERSITY OF ALBERTA
FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read
and recommend to the Faculty of Graduate Studies for
acceptance, a thesis entitled RATIONAL APPROXIMATIONS
submitted by Garry R. Jackson in partial fulfilment of
the requirements for the degree of Master of Science.

John M. Macneil

.....

Supervisor

A. J. Keeping

P. G. Sinclair

S. Hunka

Date *21. 1. 64*

ABSTRACT

This thesis is concerned with approximation by rational functions.

Basic Chebyshev concepts regarding polynomial approximation are reviewed. Theorems concerning the approximation of a continuous function by a rational function are presented along with proofs. Then basic Chebyshev concepts are employed in an attempt to outline a constructive theory of approximation by rational functions.

Some algorithms for obtaining rational interpolatory functions and rational approximations have been included.

ACKNOWLEDGEMENTS

I thank Professor J. McNamee for his invaluable assistance in the preparation of this thesis and his understanding guidance over the past two years.

I also thank Dr. D. B. Scott, Director of the University of Alberta Computing Center, for providing a stimulating atmosphere in which to work.

To all my friends who assisted me in transcribing this thesis, my sincere thanks.

TABLE OF CONTENTS

| | | Page |
|---------------|---|------|
| CHAPTER I | INTRODUCTION | 1 |
| Section 1.1 | The Nature of Approximation | 1 |
| Section 1.2 | The Beginning of Approximation and Interpolation | 3 |
| Section 1.3 | The Modern Theory of Approximation | 6 |
| Section 1.4 | Survey of the Thesis | 9 |
| CHAPTER II | POLYNOMIAL APPROXIMATION | 11 |
| Section 2.1 | Introduction | 11 |
| Section 2.2 | The Taylor Polynomial | 11 |
| Section 2.3 | The Lagrange Polynomial | 13 |
| Section 2.4 | The Newton Polynomials and Divided Differences | 16 |
| Section 2.5 | Error Estimates for the Lagrange- Newton Interpolation Polynomials | 18 |
| Section 2.6 | Chebyshev Concepts | 21 |
| CHAPTER III | RATIONAL APPROXIMATION | 31 |
| Section 3.1 | Introduction | 31 |
| Section 3.2 | Rational Approximations to Continuous Functions: Statement of the Problem | 31 |
| Section 3.3 | Achiezer's Form of Chebyshev's Theorem | 33 |
| Section 3.3.1 | A Generalization of de la Vallée Poussin's Theorem | 33 |

| | | Page |
|---------------|--|------|
| Section 3.3.2 | Achiezer's Form of Chebyshev's Uniqueness Theorem | 38 |
| Section 3.4 | Chebyshev's Theorem | 42 |
| Section 3.5 | Rational Approximations to Sets of Ordinates | 54 |
| Section 3.5.2 | Examples | 62 |
| Section 3.5.3 | Interpretation of the Examples | 68 |
| Section 3.5.4 | Methods for Obtaining ρ | 71 |
| CHAPTER IV | METHODS OF FINDING RATIONAL INTERPOLATION FORMULAE | 81 |
| Section 4.1 | Introduction | 81 |
| Section 4.2 | Some Relationships between Finite Continued Fractions and Rational Functions | 81 |
| Section 4.3 | The Inverse Divided-Difference Interpolation Formula | 85 |
| Section 4.4 | Thiele's Continued-Fraction Expansions | 89 |
| Section 4.5 | The Error Term of Thiele Expansions | 92 |
| Section 4.6 | Thacher and Tukey's Algorithm for Obtaining Rational Interpolates | 96 |
| CHAPTER V | ALGORITHMS FOR OBTAINING RATIONAL APPROXIMATIONS | 108 |
| Section 5.1 | Introduction | 108 |
| Section 5.2 | The Padé Approximants | 109 |
| Section 5.3.1 | An Extension of Remez's Second Algorithm | 113 |

| | Page |
|---|------|
| Section 5.3.2 The Loeb Weighted Minimax Algorithm | 119 |
| Section 5.3.3 Loeb's Linear Inequality Method | 121 |
| Section 5.3.4 A Differential Correction Method | 123 |
| Section 5.4 Maehly's Telescoping Procedure for Rational Approximations | 124 |
| Section 5.5 Maehly's "Direct Methods" for Fitting Rational Approximations | 130 |
| Section 5.6 Maehly's "Indirect" and "Combined" Methods of Fitting Rational Approximations | 138 |
| CHAPTER VI CONCLUSION | 149 |
| BIBLIOGRAPHY | 152 |
| APPENDIX | A1 |

CHAPTER I

INTRODUCTION

Section 1.1 The Nature of Approximation

In many of the problems which arise in numerical analysis, the mathematician is given certain information concerning a particular function that, for example, may be known to be - or assumed to be - continuous. He is then required to educe additional or improved information in a form which is appropriate for interpretation in terms of numbers.

A technique which is frequently used in such cases can be described, in general terms, as follows. A convenient set of $n+1$ coordinate functions, say $\varphi_0(x)$, $\varphi_1(x)$, ..., $\varphi_n(x)$, is first selected. Then a procedure is invented which has properties such that it will yield the desired additional information easily, and (except for inaccuracies in calculation) exactly if $f(x)$, the function to be approximated, is a member of the set, S_n - n maybe infinite, of the functions which are expressible exactly as linear combinations of the coordinate functions. Here we are assuming that an appropriate criterion for approach, such as the Chebyshev criterion (see below and Chapter 2), least squares criterion, etc.,

has been chosen. Different criteria will, in general, yield different approximations, and any particular criterion will reproduce some properties of $f(x)$ more faithfully, other less faithfully. Some loss of fidelity must be accepted in an approximation; but - as mentioned above - the selective process should arrive at $f(x)$ if $f(x)$ is in S_n .

One further practical point is recognized. In general, the first approximation to a function uses a bare minimum of information and thus is crude. The process must, in general be such that the approximation can be improved by incorporating additional information. It is also desirable that an existing approximation should be able to use additional information concerning $f(x)$, but not used by the selective process, for estimating the error of the approximation.

Obviously, it is advantageous to choose coordinate functions whose properties are simple and which are easily computed. Since polynomials are easily evaluated and since their integrals, derivatives, and products are also polynomials, the $n+1$ functions, $1, x, \dots, x^n$, which generate the algebraic polynomials of degree $\leq n$, are particularly appropriate - again, n may be infinite.

The history of polynomial, and other approximations spans two centuries and only a brief outline of developments of major importance will be given here - not necessarily in chronological order.

Section 1.2 The Beginning of Approximation and Interpolation

Interpolation may be defined as any method for obtaining numerical values of a function which uses only certain tabular values of that function.

It appears as though Gregory should be regarded as the father of interpolation based on the representation of functions by power polynomials. Although Newton's interpolation formulae using forward and backward differences first appeared in print in his "Philosophiae Naturales Principia Mathematica", Oxford, 1687 (Book III, Lemma V), they were in fact discovered by Gregory some seventeen years earlier. Briggs originated the use of tabular differences of higher than the first order for the construction and interpolation of tables in his "Arithmetica Logarithmia", London, 1624, and later in "Trigonometric Britannica", Gaudae, 1633, which he and Gellibrand wrote. Briggs appears to have preceded Gregory but he did not develop any interpolation formulae using the higher order differences. The terms "interpolatio" and "interpolare"

in their present sense were apparently introduced by Wallis in his "Arithmetica Infinitorum", Oxford, 1659.

Newton, in his "Principia" (Book III, Lemma V) gave the general notation of divided differences while the term itself appears to have been coined by DeMorgan in his "Differential Calculus", London, 1842. In addition, Newton gave us two interpolation formulae, commonly known as Sterling's and Bessel's, first with unequal, then with equal argument intervals, in "Methodus Differentialis" of the "Analysis per quantitatum, Series, Fluciones, ac Differentias...", London, 1711. It can also be shown that Gauss's interpolation formula is only a slight transformation of Sterling's (Newton's) formula (see p.66 of [18]).

A result of fundamental importance in numerical work is the Lagrange interpolation formula which was closely anticipated by Euler in his "Institutiones Calculi Differentialis", Petrograd, 1755, in which he reached his formulae by considering series of which a certain order of difference is constant. The highest polynomial, one of the fifth order, that Euler gave was deduced from Newton's forward difference formula. However, the Lagrange formulae was given explicitly by Waring in "Phil. Trans.

Roy. Soc." 69, 59 (1779) and was subsequently published by Euler in his "Opuscula Analytica", (Vol. I), Petrograd, 1783, in the form appropriate for use with equidistant data. Apparently this latter publication does not antedate Waring, but it does antedate Lagrange, who did not publish the formula that is now generally associated with his name until 1795 in "Jour. de l'École Polytechnique", 2, 274. The logical connection between Lagrange's (Waring's) and Newton's divided difference interpolation formulae was first explicitly pointed out by Gauss.

Another central difference interpolation formula, commonly known as Everett's formula, was published in 1901 by Everett in "Jour. Inst. Act." 35, 452. However, the formula was actually first derived by Laplace in his "Théorie Analytique des Probabilités", Paris. This formula, unlike other central-difference formulae, uses only even-order central differences, and, as a result, if only lower-order differences are used, requires less tabulation of differences. Another formula, usually referred to as Steffenson's interpolation formula, was credited to Everett by Steffenson who called it Everett's second formula in his "Interpolation".

It can be shown (see p.16 of [18]) that the

forward-difference, central-difference, and the Lagrangian formulae, when the abscissae are equally spaced, are really different aspects of the same process; that is, the running of a polynomial of degree $n-1$ through n points. Hence when we use these interpolation formulae, we are assuming that the interpolation polynomial and the function coincide closely enough to give an error of interpolation that lies within prescribed bounds. However, because the higher-order differences involved in the forward and backward-difference formulae are less and less closely related to the behaviour of the function near the desired point, these interpolation formulae are not as accurate in practice as other difference formulae. Thus the forward and backward-difference formulae should only be used in cases where it is impossible to use other interpolation formulae; that is, at the ends of the difference table. Even then, it is, if possible, better to extend the difference table and use a central-difference interpolation formula.

Section 1.3 The Modern Theory of Approximation

Approximation may be described as the act of replacing a known, untractable function by a tractable function having similar characteristics.

Let us, at this point, exemplify how the criterion of approximation influences the nature of the interpolatory approximation. (All the approximations considered at this point are polynomial approximations). The use of zeros of the Legendre polynomials in approximation minimizes the root mean square value of the error over the open interval $(-1, 1)$ whereas the use of the zeros of the Chebyshev polynomials minimizes the largest of the absolute values of the errors. Furthermore, the errors oscillate uniformly over $(-1, 1)$, while in the first case they tend to oscillate with increasing amplitude towards the ends of the interval. It was shown by Chebyshev in 1853 that, when the maximum-error criterion is used, the choice of the Chebyshev polynomials is the best possible one. In the series of theorems containing the above result he gave a practical and simple construction which enables one to find excellent approximations under mild restrictions. A discussion of Chebyshev concepts will be found in Chapter II.

A natural extension of polynomial approximation is approximation by rational functions; that is, by the ratio of two polynomials. Cauchy, in his "Cours d'analyse de l'École polytechnique", premier Partie, Note V, Paris, 1821, proposed an interpolation formula involving the ratio of two polynomials of different orders. Thus,

Cauchy was one of the first to investigate the possibility of interpolation and approximation by rational functions. In fact, it appears to be very difficult to find any branch of analysis upon which Cauchy has not touched. In 1846, E. Brassinne published a paper in "Journal de Mathematiques pures et appliquées", Tome XI, Paris, which indicated the variety of formulae which are included under Cauchy's ratio formula. Theile, in his "Interpolationsrechnung", B. G. Teubner, Leipzig, 1909, published his reciprocal-difference algorithm. This algorithm allows the construction, from a table of reciprocal differences, of a rational approximation in continued fraction form, which is the easiest and most efficient form for evaluation of rational functions of interpolation type. The algorithm will be given in Chapter IV.

Up to this point, we have considered mainly interpolating approximations. This type of approximation may be characterized as approximation to differentiable functions. A radically different method of approximation was initiated by Weierstrass when he proposed to consider approximations to continuous functions by a polynomial of sufficiently high order. Weierstrass published his theorem on polynomial approximation in "Sitzungsber Akad Berlin", in 1885. This famous theorem states, in effect,

that any function continuous over a closed interval can be uniformly approximated over that interval, to any prescribed tolerance, by a polynomial of sufficiently high order. Bernstein's theorem, published in 1912, is a significant improvement over Weierstrass's result, since it actually constructs a sequence of polynomials such that the approximation improves steadily as the order of the approximating polynomial increases. However, Bernstein's method converges too slowly to be of much practical value. Today there is much interest in the Chebyshev theorems and with modern high-speed digital computers we are increasingly able to find Chebyshev-approximants to continuous functions.

Section 1.4 Survey of the Thesis

The remainder of this thesis will be concerned with the following topics. Chapter II consists of a presentation of the Taylor polynomials, the Lagrange formula, the Newton polynomials and some Chebyshev concepts regarding polynomial approximations. Chapter III contains an extension of Chebyshev concepts to rational approximations, while chapter IV contains some algorithms for finding rational interpolants, and chapter V contains algorithms that give - in most cases - "best-fit" (in the

Chebyshev sense) rational approximations. Chapter VI contains a summary of what has been covered in this thesis and a discussion of some significant unsolved problems concerning rational approximations.

CHAPTER II

POLYNOMIAL APPROXIMATION

Section 2.1 Introduction

This chapter contains a discussion of the Taylor polynomials and the development and a discussion of the Lagrangian formulae, the Newton polynomials and some basic Chebyshev concepts concerning polynomial approximation. The Taylor polynomials are included because they are among the forms most widely used on today's high speed computers.

When the notation $f(x)$ is used in what follows, it is meant to represent an arbitrary continuous function defined in a finite interval $a \leq x \leq b$, and in section 2.2, $f(x)$ is differentiable any number of times at some point c contained in the interval of interest.

Section 2.2 The Taylor Polynomial

If the Taylor series of a function $f(x)$, the general term of which is

$$\frac{(x-c)^k f^{(k)}(c)}{k!},$$

is truncated after n terms, the resulting expression is a polynomial of degree $\leq n$. Using the mean-value theorem, which states that

$$f' [c + \theta(x-c)] = \frac{f(x) - f(c)}{x-c}$$

for some θ satisfying $0 < \theta < 1$, we now consider

$$(2.2.1) \quad g(\xi) = F_n(\xi) - \left(\frac{x-\xi}{x-c}\right)^{n+1} F_n(c)$$

where

$$(2.2.2) \quad F_n(\xi) = f(x) - f(\xi) - \sum_{k=1}^n \frac{(x-\xi)^k}{k!} f^{(k)}(\xi).$$

Clearly the right hand side of (2.2.1.) vanishes at $\xi=x$ and $\xi=c$. Hence by Rolle's theorem its derivative with respect to ξ must vanish for some value of ξ in the interval (x,c) . The derivative of $g(\xi)$ is

$$g'(\xi) = \frac{(n+1)(x-\xi)^n}{(x-c)^{n+1}} \left[F_n(c) - \frac{(x-c)^{n+1}}{(n+1)!} f^{(n+1)}(\xi) \right].$$

Therefore we have

$$F_n(c) = \frac{(x-c)^{n+1}}{(n+1)!} f^{(n+1)}[c+\theta(x-c)]$$

and from (2.2.2) we see that this is the difference between $f(x)$ and the Taylor polynomial and is equal to

$$R_n(x) = \frac{(x-c)^{n+1}}{(n+1)!} f^{(n+1)}[c+\theta(x-c)].$$

An evaluation of the Taylor polynomials yields the following points.

1. Within the range of convergence, the Taylor polynomial yields a uniform approximation to $f(x)$ which improves steadily as n increases.

2. The derivative of the Taylor polynomial uniformly approaches the derivative of $f(x)$ as n increases.
3. Unless we can eliminate the dependence on the derivative, the error estimate is almost unusable.
4. The Taylor polynomial is obviously not the best approximation to $f(x)$ because it relies on the behavior of the function at one point only.
5. The necessary data for the Taylor polynomial of degree n consists of $n+1$ numerical values: $f(x)$ and its first n derivatives evaluated at the single point c .

From here we will continue on to much more accurate forms of approximation - those that use information from more than one abscissa.

Section 2.3 The Lagrange Polynomial

The ordinates y_i involved in Lagrange's form of the polynomial $L_n(x) \equiv y_{0\dots n}(x)$ of degree n , which takes on the same values as a given function $f(x)$ for the $n+1$ distinct abscissae x_0, x_1, \dots, x_n , are displayed explicitly.

We can write $L_n(x)$ directly in the required form.

$$(2.3.1) \quad L_n(x) = l_0(x)f(x_0) + \dots + l_n(x)f(x_n)$$

$$\equiv \sum_{k=0}^n l_k(x)f(x_k),$$

where the $l_k(x)$ are polynomials of degree n , to be determined by the requirement that $L_n(x_i) = f(x_i)$, $i = 0, 1, \dots, n$. Now the expression (2.3.1) will take on the value $f(x_i)$ at $x = x_i$ if

$$l_i(x_j) = \delta_{ij} \quad i = 0, 1, \dots, n, \quad j = 0, 1, \dots, n,$$

where δ_{ij} is the Kronecker delta.

Since $l_i(x)$ is to be a polynomial of degree n which vanishes for $x = x_0, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$, it must follow that

$$(2.3.2) \quad l_i(x) = c_i \prod_{\substack{j=0 \\ j \neq i}}^n (x - x_j)$$

where c_i is a constant. The final requirement that $l_i(x_i) = 1$ gives us the following expression for c_i

$$(2.3.3) \quad c_i = \prod_{\substack{j=0 \\ j \neq i}}^n (x_i - x_j)^{-1}$$

and we obtain the desired influence coefficients $l_i(x)$ by

introducing (2.3.3) into (2.3.2).

We can put this result in a more compact form by using the notation

$$(2.3.4) \quad P_{n+1}(x) = \prod_{j=0}^n (x-x_j).$$

The derivative of $P_{n+1}(x)$ is expressible as the sum of $n+1$ terms in each of which one of the factors of $P_{n+1}(x)$ is deleted. Thus if we set $x = x_i$ in this expression, we arrive at

$$P'_{n+1}(x_i) = \prod_{\substack{j=0 \\ j \neq i}}^n (x_i - x_j) = 1/c_i$$

Hence we can write (2.3.1) in the form

$$L_n(x) = \sum_{k=0}^n \frac{P_{n+1}(x)f(x_k)}{(x-x_k)P'_{n+1}(x_k)}.$$

It is obvious that the coefficients of x^n in $l_i(x)$ is $1/P'_{n+1}(x_i)$ and that the coefficient of x^n in $L_n(x)$ is then $\sum_{i=0}^n f(x_i)/P'_{n+1}(x_i)$.

Section 2.4 The Newton Polynomials and Divided Differences

Let $L_i(x)$ denote the polynomial of degree i which passes through the points $(x_0, y_0), (x_1, y_1), \dots, (x_i, y_i)$. Then $L_n(x)$ can be written as

$$(2.4.1) \quad L_n(x) = y_0 + \sum_{i=1}^n (L_i(x) - L_{i-1}(x)).$$

Now by definition $L_i(x) - L_{i-1}(x)$ vanishes at x_0, x_1, \dots, x_{i-1} and so does

$$(2.4.2) \quad \prod_{k=0}^{i-1} (x - x_k).$$

Hence the $L_i(x) - L_{i-1}(x)$ and the product can differ only by a constant which we can find by comparing the coefficients of x^i in the product and in $L_i(x) - L_{i-1}(x)$. The coefficient of x^i in $L_i(x)$ is

$$(2.4.3) \quad \sum_{k=0}^i \frac{y_k}{P'_{i+1}(x_k)} = y(x_0, \dots, x_i) \quad .$$

Hence

$$\begin{aligned} (2.4.4) \quad L_n(x) &= y_0 + \sum_{i=1}^n [L_i(x) - L_{i-1}(x)] \\ &= y_0 + (x - x_0)y(x_0, x_1) + (x - x_0)(x - x_1)y(x_0, x_1, x_2) \end{aligned}$$

$$+ \dots + (x - x_0)(x - x_1) \dots (x - x_{n-1})y(x_0, x_1, \dots, x_n).$$

Now let us attempt to find the $y(x_0, \dots, x_1)$ by recursion.

To do this let us examine the linear polynomial defined by (x_0, y_0) and (x_1, y_1) . Using (2.4.4) we have

$$(2.4.5) \quad L_1(x) = y_0 + (x - x_0)y(x_0, x_1)$$

$$(2.4.6) \quad = y_1 + (x - x_1)y(x_1, x_0).$$

If we put $x = x_1$ in (2.4.5) and $x = x_0$ in (2.4.6) we get

$$\frac{y_1 - y_0}{x_1 - x_0} = y(x_0, x_1) = y(x_1, x_0).$$

Let us now examine the quadratic polynomial defined by $(x_1, y_1), (x_2, y_2)$ and (x_0, y_0) .

$$(2.4.7) \quad L_2(x) = y_1 + (x - x_1)y(x_1, x_2) + (x - x_1)(x - x_2)y(x_0, x_1, x_2).$$

Putting $x = x_0$ in (2.4.6) and (2.4.7) gives on subtraction

$$(x_0 - x_1)y(x_0, x_1) = (x_0 - x_1)y(x_1, x_2) + (x_0 - x_1)(x_0 - x_2)y(x_0, x_1, x_2)$$

so that

$$y(x_0, x_1, x_2) = \frac{y(x_1, x_2) - y(x_0, x_1)}{x_2 - x_0}.$$

If we proceed in this way we find that

$$y(x_0, x_1, \dots, x_{i+1}) = \frac{y(x_1, x_2, \dots, x_{i+1}) - y(x_0, x_1, \dots, x_i)}{x_{i+1} - x_0}$$

or more generally

$$y(x_i, x_{i+1}, \dots, x_{i+k+1}) = \frac{y(x_{i+1}, x_{i+2}, \dots, x_{i+k+1}) - y(x_i, x_{i+1}, \dots, x_{i+k})}{x_{i+k+1} - x_i}.$$

These expressions are called "divided differences". The polynomial resulting from use of this method of approximation is the same polynomial as given by the Lagrange Method and has the advantage that when we wish to find an approximating polynomial of higher degree we need only extend the difference table, whereas if we used the Lagrange form we would have to modify the complete system.

Section 2.5 Error Estimates for the Lagrange-Newton

Interpolation Polynomials

Let us emphasize the construction of the approximation by recalling the notation $y_i = f(x_i)$, $i=0, 1, \dots, n$. The polynomial constructed from the ordinates yields at best an approximation to the value of $f(X)$ at any point X not equal to any of the x_i .

If we denote the value of $f(X)$ by Y and the value of $L_n(X)$ by Y_L we then need the measure of the deviation $Y - Y_L$.

Let $L_{n+1}(x)$ be the Lagrange - Newton polynomial constructed from the points $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n), (X, Y)$. Then

$$(2.5.1) \quad L_{n+1}(x) = y_0 + (x - x_0)y(x_0, x_1) + \dots + (x - x_0)(x - x_1) \dots$$

$$(x - x_{n-1})y(x_0, \dots, x_n) + (x - x_0)(x - x_1) \dots$$

$$(x - x_n)y(x_0, x_1, \dots, x_n, X),$$

and

$$L_{n+1}(X) = Y,$$

and

$$L_n(X) = Y_L.$$

If we subtract (2.4.4) from (2.5.1) we get

$$L_{n+1}(X) - L_n(X) = (X - x_0)(X - x_1) \dots (X - x_n)y(x_0, x_1, \dots, x_n, X).$$

However this is not of much use in its present form because $y(x_0, x_1, \dots, x_n, X)$ uses the very ordinate Y which we are trying to approximate. Hence let us try to put the error

estimate in a different form. To do this we make the following new assumption; let the $(n+1)$ th derivative of $f(x)$ exist in some interval $[a,b]$ containing the x_1 and X . Now let us consider $f(x) - L_{n+1}(x)$ which vanishes at the points x_0, x_1, \dots, x_n, X . Hence, by repeated application of Rolle's theorem, the first, second, \dots $(n+1)$ th derivatives of the difference vanish $n+1, n, \dots, 1$ times respectively in $[a,b]$.

Let ξ be the point where the $(n+1)$ th derivative vanishes.

Then

$$\frac{d^{n+1}}{dx^{n+1}} [f(x) - L_{n+1}(x)]_{x=\xi} = 0.$$

That is

$$(2.5.2) \quad f^{n+1}(\xi) = (n+1)! \, y(x_0, x_1, \dots, x_n, X),$$

and our error estimate becomes

$$Y - Y_L = \frac{(X - x_0)(X - x_1) \dots (X - x_n) f^{n+1}(\xi)}{(n+1)!}$$

This leads us to the consideration of the question - what is the best polynomial approximation? - which is considered in the following section.

Section 2.6 Chebyshev Concepts

Chebyshev posed the following question: if the degree, n , of the approximating polynomial is fixed in advance, can we extract from the set of all polynomials of degree $\leq n$ that polynomial, $p(x)$, which approximates $f(x)$, $[f(x) \in C[a,b]]$, most closely in the sense that $\max |f(x)-p(x)|$ is least in $[a,b]$? This may be expressed in a more mathematical form. If $p(x) = c_0 + c_1x + \dots + c_nx^n$, then the greatest of the deviations $|f(x)-p(x)|$, $x \in [a,b]$, is a positive function of the coefficients c_i , which we may denote by $\rho(c_i)$. If we allow the c_i to vary continuously over all real values then $\rho(c_i)$ has a least value and we seek the polynomial which yields the least value.

We shall find it convenient in the sequel to abbreviate this statement of the question and speak of the polynomial of approximation ρ . We shall denote the least value of ρ by $\min \rho$. Let us use the notation:

ordinates y_0, y_1, \dots, y_{n+1}

abscissae x_0, x_1, \dots, x_{n+1} with the ordering $x_{i+1} > x_i$.

To fix the concept let us take a special case of this problem first: a linear polynomial approximating three points.

Let the ordinates of the approximating line be y_0-u_0 , y_1-u_1 , and y_2-u_2 . The straight line with the ordinates y_0-u_0 and y_1-u_1 at x_0 and x_1 is

$$(2.6.1) \quad y(x) = \frac{x-x_1}{x_0-x_1} (y_0-u_0) + \frac{x-x_0}{x_1-x_0} (y_1-u_1).$$

where the coefficients of the ordinates on the right are the Lagrangian polynomials. This line passes through (x_2, y_2-u_2) , and hence

$$y_2-u_2 = \frac{x_2-x_1}{x_0-x_1} (y_0-u_0) + \frac{x_2-x_0}{x_1-x_0} (y_1-u_1).$$

Now multiplying through by (x_1-x_0) we have

$$(2.6.2) \quad (x_2-x_1)(y_0-u_0) - (x_2-x_0)(y_1-u_1) + (x_1-x_0)(y_2-u_2) = 0$$

or

$$(2.6.3) \quad u_0 q_0 - u_1 q_1 + u_2 q_2 = y_0 q_0 - y_1 q_1 + y_2 q_2$$

where the q_i are the positive differences of the abscissæ.

It is convenient to express the u_i in terms of the maximum deviation which occurs at one or more of the x_i . Thus we write $u_i = \rho v_i$ where $|v_i| \leq 1$ and at least one of the

$v_i = \pm 1$. We can regard the v_i as arbitrary and write (2.6.3) as an equation to determine ρ . We get

$$\rho = \frac{|y_0 q_0 - y_1 q_1 + y_2 q_2|}{|v_0 q_0 - v_1 q_1 + v_2 q_2|},$$

where the y_i and q_i are fixed and $q_i > 0$. Only the v_i are variable and hence are at our disposal. Clearly the denominator is greatest and ρ is least when the v_i alternate in sign and each has the greatest admissible value, unity. Hence

$$\min \rho = \frac{|y_0 q_0 - y_1 q_1 + y_2 q_2|}{q_0 + q_1 + q_2}.$$

This shows that the best approximation for this special case, in the Chebyshev sense, is the one which has the maximum deviation occurring at three points with alternating sign.

We shall now consider the general case of $n+2$ ordinates, $y_i, i=0, 1, \dots, n+1$. Let us denote the ordinates of the approximating polynomial, which is of degree n , by $y_i - u_i$. Now the polynomial of degree n which passes through the $n+1$ points $(x_i, y_i - u_i), i=0, 1, \dots, n$ is

$$(2.6.4) \quad L_n(x) = \sum_{i=0}^n (y_i - u_i) \ell_i(x).$$

Where

$$(2.6.5) \quad \ell_i(x) = \frac{\prod_{\substack{k=0 \\ k \neq i}}^n (x - x_k)}{\prod_{\substack{k=0 \\ k \neq i}}^n (x_i - x_k)}$$

(the Lagrangian coefficients).

Before we continue, let us note the following two points.

1. The numerator of $\ell_i(x)$ is positive for $x > x_n$.
2. The sign of the denominator is $(-1)^{n-i}$. (In particular, the sign of $\ell_i(x_{n+1})$ is $(-1)^{n-i}$, $i=0,1,\dots,n$, that is, the $\ell_i(x_{n+1})$ alternate in sign.)

On putting $x = x_{n+1}$ in (2.6.4) we have

$$(2.6.6) \quad y_{n+1} - u_{n+1} = \sum_{i=0}^n (y_i - u_i) \ell_i(x_{n+1}).$$

If we now multiply (2.6.6) by the position factor

$$q_{n+1} = \prod_{\substack{i,j=0 \\ i>j}}^n (x_i - x_j)$$

we can write (2.6.6) in the form

(2.6.7)

$$(y_{n+1} - u_{n+1})q_{n+1} = \sum_{i=0}^n (-1)^{i-n} (y_i - u_i)q_i$$

where the q_i are the product of all the absolute value differences

$$|x_j - x_k|, j, k \neq i, j > k, j, k = 0, 1, \dots, n+1.$$

Writing (2.6.7) in the symmetric form

(2.6.8)

$$\sum_{i=0}^n (-1)^{i-n} (y_i - u_i)q_i - (y_{n+1} - u_{n+1})q_{n+1} = 0$$

and $u_i = \rho v_i, |v_i| \leq 1$, we have

(2.6.9)

$$\rho = \frac{|y_0 q_0 - y_1 q_1 + \dots + (-1)^n y_n q_n + y_{n+1} q_{n+1}|}{|v_0 q_0 - v_1 q_1 + \dots + (-1)^n v_n q_n + v_{n+1} q_{n+1}|}.$$

It is now obvious that the minimum value of ρ occurs when $|v_i| = 1$ and the signs of the v_i alternate since the y_i and the q_i are fixed and are not at our disposal.

We can draw two very useful conclusions from

this result. Now the y_i and q_i are fixed and the q_i are positive. Hence only the v_i are at our disposal and $|v_i| \leq 1$. Thus we have

(2.6.10)

$$\min \rho = \left| \frac{y_0 q_0 - y_1 q_1 + \dots + (-1)^n y_n q_n + y_{n+1} q_{n+1}}{q_0 + q_1 + \dots + q_n + q_{n+1}} \right|.$$

The second conclusion that we shall draw is much more subtle. Let us for the moment go back to the u_i notation and write (2.6.8) in the form

(2.6.11)

$$y_0 q_0 - y_1 q_1 + \dots + y_{n+1} q_{n+1} = u_0 q_0 - u_1 q_1 + \dots + u_{n+1} q_{n+1}.$$

We can also write

(2.6.12)

$$y_i = (y_i - u_i) + u_i.$$

Now the preceding analysis shows that we can fit a polynomial of degree $\leq n$ through the $n+2$ points $(x_i, y_i - u_i)$, $i=0, 1, \dots, n+1$, provided only the u_i are related to the y_i by (2.6.11) or, what is equivalent, provided that ρ is given by (2.6.9). Hence from (2.6.12) we see that the problem of approximating the y_i by a polynomial of degree $\leq n$ is equivalent to the

problem of approximating the u_i by a polynomial of degree $\leq n$.

Let us denote the new approximation by ρ' .

Having done so, we get

(2.6.13)

$$\min \rho' = \frac{|u_0 q_0 - u_1 q_1 + \dots + (-1)^n u_n q_n + u_{n+1} q_{n+1}|}{q_0 + q_1 + \dots + q_n + q_{n+1}}.$$

In particular, if the u_i alternate in sign and if we use the notation $r_i = |u_i|$ we have

(2.6.14)

$$\min \rho' = \frac{r_0 q_0 + r_1 q_1 + \dots + r_{n+1} q_{n+1}}{q_0 + q_1 + \dots + q_{n+1}}.$$

Since the problems are equivalent we have

$$\min \rho' = \min \rho.$$

From (2.6.14) we see that, in effect, $\min \rho'$ lies between the greatest and the least values of r_i , that is, it is a weighted average.

These two conclusions lead us to the following theorems.

Theorem (2.6.1) There exists a unique polynomial of degree $\leq n$ which furnishes the closest approximation to a set of $n+2$ distinct ordinates. The deviations between the ordinates of the approximating polynomial and the prescribed ordinates alternate in sign and are of equal magnitude.

Theorem (2.6.2) If a polynomial of degree $\leq n$ furnishes deviations u_i (from $n+2$ prescribed ordinates) of alternating signs, but unequal magnitude, then the closest approximation to the prescribed ordinates lies between the greatest and the least of the absolute value of the u_i and is the weighted average of the u_i ; the weights being the q_i .

We can now return to the original problem by generalizing theorem (2.6.1) in two steps. First we consider the best approximation to a set of m ordinates ($m > n+2$). Then we consider the case in which the ordinates may have any abscissae in a continuous strip $[a, b]$. The preceding work enables us to formulate the following theorem.

Theorem (2.6.3) The best polynomial approximation of degree $\leq n$ to a finite set of m ($m > n+2$) points is that one corresponding to the greatest of the $\binom{m}{n+2}$ values of $\min \rho$ at $n+2$ points selected arbitrarily from m points.

Proof:

Let us use the notation

σ for the greatest of the values of $\min \rho$,

E for the set of $n+2$ points which furnishes σ ,

y_1 for the set of ordinates of these points.

If the absolute value of the deviation u , at any point p (of the set of points) not belonging to E , is greater than σ , then let E' be the set of points obtained by deleting one of the points of E and adding p , the deletion being made in such a way that deviations of the points of E' alternate in sign. Write $u_p = \sigma + \tau$, $\tau > 0$. Then from theorem 2

$$\min_{E'} \rho = \frac{(\sigma + \tau)q'_0 + \sigma q'_1 + \dots + \sigma q'_{n+1}}{q'_0 + q'_1 + \dots + q'_{n+1}} > \sigma ,$$

which is a contradiction that proves the theorem.

We also have the following corollary to theorem (2.6.3).

Corollary: If $f(x)$ is continuous in $[a, b]$, the closest polynomial approximation of degree $\leq n$ is that which corresponds to the greatest of the values of $\min \rho$ on any set of $n+2$ points of $[a, b]$.

Proof:

The best approximation over a set of $n+2$ points

in $[a,b]$ is given by (2.6.10). This formula demonstrates that ρ is a continuous function of the x_1 in the set in $[a,b]$. Now this function admits a maximum ρ and attains it for a particular set E in $[a,b]$. Hence we can prove the corollary by the same argument as that employed to prove theorem (2.6.3).

CHAPTER III

RATIONAL APPROXIMATION

Section 3.1 Introduction

In the last chapter we observed that there is always a polynomial approximation, of degree $\leq n$, of best fit (in the Chebyshev sense) to a set of $n + 2$ points. In this chapter we will investigate rational approximations to continuous functions. We will then proceed to the difficult problem: Using a similar line of attack to that of Chapter II section 6 can we find a \min_p for a rational approximation of the form $p_n(x)/q_m(x)$, where $p_n(x)$ is a polynomial of degree $\leq n$ and $q_m(x)$ is a polynomial of degree $\leq m$, to a given set of $n + m + 2$ points? Also, does a best rational approximation always exist?

Section 3.2 Rational Approximations to Continuous Functions:Statement of the Problem

Chebyshev's formulation of a class of problems of this nature is admirable and lucid and we follow him here.

It is required to determine a function $y(x)$ involving $n + 1$ arbitrary parameters p_1, p_2, \dots, p_{n+1} such that the greatest deviation $|y(x) - f(x)|$ between $y(x)$ and the given function $f(x)$ in the interval $-a \leq x \leq a$ is least. More precisely, we require

$$\min_{p_i} \max_{-a \leq x \leq a} |y(x) - f(x)|$$

This formulation embraces the following problems.

1. To minimize with respect to the p_i

(3.2.1)

$$r(x) = | p_0 x^{n-v} + p_1 x^{n-v-1} + \dots + p_{n-v} - f(x) |$$

$$\text{in } -a \leq x \leq a.$$

2. To minimize with respect to the p_i

(3.2.2)

$$r(x) = \left| \frac{p_0 x^{n-v} + p_1 x^{n-v-1} + \dots + p_{n-v}}{q_0 x^{m-\mu} + q_1 x^{m-\mu-1} + \dots + q_{m-\mu}} - f(x) \right|$$

$$\text{in } -a \leq x \leq a. \text{ The } q_i \text{ are given.}$$

3. To minimize with respect to the p_i and q_i

(3.2.3)

$$r(x) = \left| \frac{p_0 x^{n-v} + p_1 x^{n-v-1} + \dots + p_{n-v}}{q_0 x^{m-\mu} + q_1 x^{m-\mu-1} + \dots + q_{m-\mu}} - f(x) \right|$$

$$\text{in } -a \leq x \leq a.$$

This is sufficient generality for our present purpose. Problem 1 is a special case of problems 2 and 3. Problem 2 has a fixed denominator in the rational fraction and we shall not consider it further. Problem 3 embraces problem 1 and problem 2. Here we shall consider only problem 3.

One of the fundamental theorems concerning

problem 3 is usually quoted in the form given by Achiezer; however, a different statement of the theorem was given by Chebyshev, which is in some respects more lucid. Moreover, Chebyshev's proof of the theorem provides a good understanding of the theory.

We will consider first Achiezer's form of the theorem and then Chebyshev's. The two forms have different contents - in Achiezer's, the number of extrema for a polynomial pair of orders $n - \nu$, $m - \mu$ is defined to be $\geq m + n + 2 - \min(\nu, \mu)$, while in Chebyshev's the number of extrema is fixed and the theorem gives information about the greatest possible orders of the polynomial pair.

Section 3.3 Achiezer's Form of Chebyshev's Theorem

We will first give a number of theorems preparatory to theorem 3.3.2 which is our final result.

Section 3.3.1 A Generalization of de la Vallée Poussin's Theorem

Before stating the generalization let us recall to mind de la Vallée Poussin's theorem which may be stated thus.

Let y_i , $i = 0, 1, \dots, n+1$, be a set of $n+2$

ordinates at the abscissae x_1 and let $y(x)$ be a polynomial of degree $\leq n$. Let

$$r_1 = y(x_1) - y_1 \quad i = 0, 1, \dots, n+1$$

and let the r_1 alternate in sign. Then the polynomial of best approximation (of order $\leq n$) on this set of points has a deviation which lies between the greatest and least of the absolute values of the r_1 .

The significant point here is that the r_1 must alternate in sign. The lower bound is useful, the upper bound less so since in approximating to a continuous function we must consider all possible sets of $n + 2$ points.

With this preamble we proceed to the generalization. We use the notation

(3.3.1.1)

$$\begin{aligned} Q(x) &= s(x) \frac{p_0 x^{n-v} + p_1 x^{n-v-1} + \dots + p_{n-v}}{q_0 x^{m-\mu} + q_1 x^{m-\mu-1} + \dots + q_{m-\mu}} \\ &= s(x) \frac{p(x)}{q(x)} . \end{aligned}$$

Here $s(x)$ is a continuous function which never vanishes in the closed interval $[-a, a]$, m and n are given, and $0 \leq \mu \leq m$, $0 \leq v \leq n$. The point of including the parameters μ and v is that we wish to consider all

polynomials $p(x)$ of degree $\leq n$ and all polynomials $q(x)$ of degree $\leq m$. No special notation is needed for this since the coefficients of $p(x)$ and $q(x)$ - in particular, the leading coefficients - may have any finite values, zero or non-zero. It is, however, convenient in the analysis which follows to indicate explicitly the degrees of a particular $p(x)$ and particular $q(x)$ by using the parameters μ and ν .

We assume that $p(x)$ and $q(x)$ have no common divisor and that $q(x)$ does not vanish in $[-a, a]$.

Let $f(x)$ be continuous in $[-a, a]$ and let the maximum absolute deviation between $f(x)$ and $Q(x)$ in $[-a, a]$ be ρ_Q ; that is,

$$\rho_Q = \max_{-a \leq x \leq a} |Q(x) - f(x)|.$$

We can now state

Theorem 3.3.1.1: A generalization of de la Vallée Poussin's theorem.

Let $Q(x) = s(x) \frac{p(x)}{q(x)}$ remain finite in $[-a, a]$ and let the deviation

$$Q(x_i) - f(x_i)$$

at the consecutive points x_1, x_2, \dots, x_N of $[-a, a]$

assume the values

$$\lambda_1, -\lambda_2, \dots, (-1)^{N-1} \lambda_N$$

where the λ_i are positive and where $N = m + n + 2 - \min(\mu, \nu)$.

Then $\rho \geq \min \lambda_i$ $i = 1, 2, \dots, N$.

Proof

$$\text{Let } R(x) = s(x) \frac{p_0' x^n + p_1' x^{n-1} + \dots + p_n'}{q_0' x^m + q_1' x^{m-1} + \dots + q_m'}$$

be such that

$$\rho_R < \min \lambda_i.$$

Then we can prove a contradiction. For writing,

$$\Delta(x) = R(x) - Q(x) = -[Q(x) - f(x)] + [R(x) - f(x)]$$

we see that the sequence of numbers

$$\Delta(x_1), \Delta(x_2), \dots, \Delta(x_N)$$

are different from zero and alternate in sign. Since $\Delta(x)$ is continuous in $[-a, a]$, it follows that $\Delta(x)$ has at least

$$N-1 = m + n + 1 - d \quad (d = \min(\mu, \nu))$$

zeros in $[-a, a]$. But that is impossible since

$$\Delta(x) = s(x) \frac{p'(x) q(x) - p(x) q'(x)}{q(x) q'(x)}$$

and the degree of the numerator is at most $m + n - d$ (at most, since we admit that the leading coefficients of $p'(x)$ and $q'(x)$ may be zero). This contradiction proves our theorem.

Corollary to Theorem 3.3.1.1. At this point we recognize that we cannot do better if we can find a $Q(x)$ whose deviation from $f(x)$ at a set of points x_i , $i = 1, 2, \dots, N$, are $\lambda, -\lambda, \dots, (-1)^{N-1} \lambda$. For Theorem 3.3.1.1 assures us that no rational function of the form $Q(x)$ can have a smaller maximum deviation. For practical purposes this is sufficient; but we can go further and show that the $Q(x)$ of minimum deviation is unique.

We are assuming existence here. For completeness we state:

Theorem 3.3.1.2: The Existence Theorem. Among the functions $Q(x)$ there exists at least one function for which ρ_Q is a minimum.

However $Q(x)$ includes the polynomials of degree $n-v$. Hence the existence of best rational approximations with a polynomial denomination is not assured by the above.

We now proceed to the final theorem of this section.

Section 3.3.2 Achiezer's Form of Chebyshev's Uniqueness

Theorem

The function $Q(x)$ of the form

$$Q(x) = s(x) \frac{p_0 x^{n-\nu} + p_1 x^{n-\nu-1} + \dots + p_{n-\nu}}{q_0 x^{m-\mu} + q_1 x^{m-\mu-1} + \dots + q_{m-\mu}}$$

which deviates less in $[-a, a]$ from a given continuous function $f(x)$ than any other function of the same form is completely characterized by the following property. The difference $Q(x) - f(x)$ assumes the maximum values $\pm \rho_Q$ not less than $m + n + 2 - d$ times in the interval $[-a, a]$, positive and negative deviations occurring alternately.

In the statement of this theorem, we have supposed that $p(x)/q(x)$ is irreducible - that is, no linear factors of the numerator correspond to the linear factors of the denominator. The proof of this theorem is quite delicate. It seems simpler to prove the theorem first for a polynomial and then modify the proof to cover the rational function $Q(x)$. The proof for the polynomial case is due to de la Vallée Poussin.

The theorem for a polynomial $p(x)$ of degree $\leq n$ amounts to the assertion that the minimum polynomial $p(x)$

has the property that $r(x) = p(x) - f(x)$ achieves its greatest value, $\pm \rho$, at least $n + 2$ times in the interval $[-a, a]$, positive and negative deviations occurring alternately. Let x_1, x_2, \dots, x_N be the abscissae of the points at which $r(x) = \pm \rho$. We prove the theorem by showing that $N \leq n + 1$ involves a contradiction. Since $r(x)$ is continuous, we can find a small interval δ_i enclosing each x_i such that $r(x)$ is one-signed in each interval δ_i .

First let us notice that if all of the maximum deviations are one-signed, we can achieve a better approximation by adding to $p(x)$ a small constant of the same sign as $-r(x)$.

Suppose then that the sequence of maximum deviations comports $k \leq n$ groups of consecutive terms of contrary sign. Let a change of sign occur between the intervals δ_i and δ_{i+1} . Clearly δ_i and δ_{i+1} are not contiguous (since $r(x)$ does not vanish in either) and we may choose a point ξ between δ_i and δ_{i+1} with the terminal points of the intervals being excluded. (See figure 1)

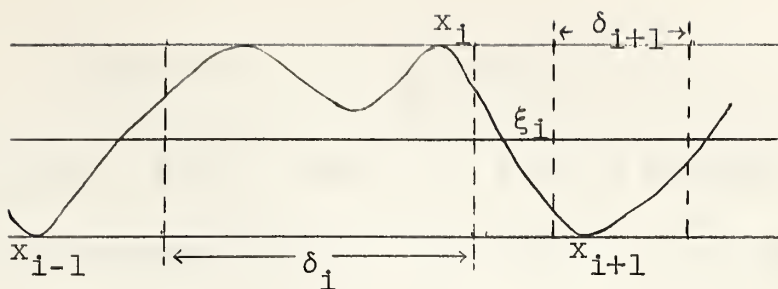
Graph of $r(x) = p(x) - f(x)$

Figure 1

Now let $\xi_I, \xi_{II}, \dots, \xi_k$ be the set of points so chosen corresponding to the k consecutive groups. Let

$$\phi(x) = [-\operatorname{sgn}(r(x_i))] \prod_{i=I}^k (\xi_i - x),$$

This follows from the fact that we want $\operatorname{sgn}[\phi(x_i)] = -\operatorname{sgn}[r(x_i)]$. Then the polynomial $\phi(x)$ has the same sign as $-r(x)$ in each of the intervals δ_i and does not vanish there. Since $k \leq n$, $\phi(x)$ is a polynomial of degree $\leq n$ and

$$p(x) + \epsilon \phi(x) \quad (\epsilon \text{ small and positive})$$

is a polynomial of degree $\leq n$ which deviates less from $f(x)$ in each interval δ_i than does $p(x)$. Now, if we can show that it deviates less in the remainder of the interval $[-a, a]$, we shall have proved our contradiction.

To complete the proof: let $\rho' < \rho$ be the bound of $|r(x)|$ in the intervals of $[-a, a]$, exterior to the intervals δ_i . Let ϵ be chosen so that

$$|\epsilon\phi| < \rho - \rho' \quad \text{in } [-a, a] \quad ;$$

then $|f(x) - p(x) - \epsilon\phi(x)| \leq |r(x)| + |\epsilon\phi(x)|$ and

for x not included in δ_i , $i = I, II, \dots, k$ we have

$$|f(x) - p(x) - \epsilon\phi(x)| < \rho' + (\rho - \rho') = \rho$$

or $|f(x) - p(x) - \epsilon\phi(x)| < \rho$.

This is a contradiction because we assumed that $p(x)$ was the best approximation with $\rho = \min \rho$.

The modification of the proof to deal with a rational function $Q(x)$ requires some minor changes. We now achieve a contradiction by assuming that the number of maximum deviations of $r(x) = Q(x) - f(x)$ is $N \leq m + n + 1 - d$ ($d = \min(\mu, \nu)$). The polynomial $\phi(x)$ is chosen as before but its degree, k , is now determined by $k \leq N - 1$; that is, $k \leq m + n - d$. We now choose polynomials $a(x)$ and $b(x)$ such that

$$\phi(x) = q(x) a(x) + p(x) b(x),$$

where the polynomial $a(x)$ is of degree $\leq n$ and $b(x)$ is of degree $\leq m$.

Now let us consider the comparison function $Q'(x)$, instead of $Q(x)$. We have

$$Q'(x) = Q(x) + s(x) \frac{\epsilon\phi(x)}{q(x)[q(x) - \epsilon b(x)]} \quad .$$

The reasoning applied in the polynomial case goes through unchanged if we add the extra condition that ϵ be chosen such that:

$$q(x) - \epsilon b(x)$$

has the same sign as $q(x)$ in $[-a, a]$.

The only point remaining to be established is that $Q'(x)$ is of the same form as $Q(x)$. But

$$\begin{aligned} Q'(x) &= s(x) \left\{ \frac{p(x)}{q(x)} + \frac{\epsilon[q(x)a(x) + p(x)b(x)]}{q(x)[q(x) - \epsilon b(x)]} \right\} \\ &= s(x) \left\{ \frac{p(x)[q(x) - \epsilon b(x)] + \epsilon[q(x)a(x) + p(x)b(x)]}{q(x)[q(x) - \epsilon b(x)]} \right\} \\ &= s(x) \left\{ \frac{p(x) + \epsilon a(x)}{q(x) - \epsilon b(x)} \right\} \end{aligned}$$

which is of the same form as $Q(x)$. Thus we reach the conclusion that $Q'(x)$ has a smaller deviation than $Q(x)$ which contradicts our original assumption.

Section 3.4 Chebyshev's Theorem

We will use the notation of (3.3.1.1), with $q_m = 1$, for $Q(x)$ in what follows.

Chebyshev's theorem can be stated as follows:

If the difference

$$Q(x) - f(x) = r(x)$$

achieves its greatest absolute value L less than $m + n + 2$ times in $[-a, a]$, then

$$p_0 = p_1 = \dots = p_{d-1} = 0$$

$$q_0 = q_1 = \dots = q_{d-1} = 0$$

where the number of extrema in $[-a, a]$ is $m + n + 2 - d$.

It is important to note that part of Chebyshev's analysis is restricted by the assumption that $f(x)$ (the function to be approximated) is differentiable. His reasoning was - in part - based on the fact that for a differentiable $f(x)$, the extrema are the common zeros of

$$r^2(x) - L^2 = 0$$

$$\text{and} \quad (x+a)(x-a) r'(x) = 0$$

in $[-a, a]$.

For $f(x)$ differentiable, there is no theoretical difficulty when the number of extrema is $\geq m + n + 2$. For the two equations above imply at least $m + n + 2$ equations for the determination of the $n + 1$ quantities p_i , $i = 0, \dots, n$, and the m quantities q_i , $i = 0, \dots, m-1$, and the deviation L . There are in principle enough equations for the solution of the problem.

Let us now consider the case where the number of extrema is less than $m + n + 2$. Although this case is more difficult, it is possible to show that it is soluble in principle. Supposing that the number of extrema is σ ($< m + n + 2$), we shall presently set up a system of $m + n + 1$ equations. When a solution exists, it is possible to obtain from these - by a process of elimination - another set of $m + n + 2 - \sigma$ equations for the determination of the $m + n + 2 + \sigma$ quantities p_i , q_i , L , x_j , $j = 1, \dots, \sigma$. Putting $x = x_j$ in both of the equations

$$r^2(x) - L^2 = 0$$

$$(x+a)(x-a) r'(x) = 0$$

we have 2σ more equations, ie. $m + n + 2 + \sigma$ in all, which is sufficient.

Proof of the Theorem

In the proof, we shall consider only the case $\sigma \leq m + n + 1$. We will use the following notation:

$$\begin{aligned} Q(x) &= \frac{p_0 x^n + p_1 x^{n-1} + \dots + p_n}{q_0 x^m + q_1 x^{m-1} + \dots + q_{m-1} x + 1} \\ &= \frac{\sum_{i=0}^n p_{n-i} x^i}{1 + \sum_{i=0}^{m-1} q_{m-1-i} x^{i+1}}, \end{aligned}$$

x_i , $i = 1, \dots, \sigma$, for the extrema of

$$r(x) = Q(x) - f(x)$$

and

$$L = \max_{-a \leq x \leq a} |r(x)|.$$

We need a number of preliminary lemmas.

Lemma 1. The quantity L is not reduced to its smallest value if the system of equations

$$(3.4.1) \quad \sum_{j=1}^{\sigma} f_{ij} \lambda_j = 0$$

where

$$\begin{aligned} f_{ij} &= \frac{\partial r(x_j)}{\partial p_{i-1}} \quad , \quad i = 1, \dots, n+1 \\ &= \frac{\partial r(x_j)}{\partial q_{i-n-2}} \quad , \quad i = n+2, \dots, n+m+1, \end{aligned}$$

has no solution other than $\lambda_j = 0$, $j = 1, \dots, \sigma$.

The proof of Lemma 1 requires two further lemmas, 1.1 and 1.2 below. However, let us note in passing that if a solution exists in which one or more of the λ_i are non-zero, then we can use σ of the equations to yield a determinantal equation for the p_i , q_i , x_i , and L . This equation, together with the remaining $m + n + 1 - \sigma$ equations constitutes a system of

$$(n + m + 1 - \sigma) + 1 = n + m + 2 - \sigma$$

equations as noted above.

Lemma 1.1. If the system (3.4.1) has no solution other than $\lambda_j = 0$, $j = 1, \dots, \sigma$, then we can find a set of numbers η_i , $i = 1, \dots, n+m+1$, such that the σ equations

$$(3.4.2) \quad \sum_{i=1}^{n+m+1} f_{ji} \eta_i = r(x_j) \quad \text{are satisfied.}$$

Proof of Lemma 1.1

There are two cases to consider:

- 1) $\sigma = m + n + 1$. Then the matrix of the f_{ij} is square and of rank $m + n + 1$. Hence the equations (3.4.2) certainly have a non-zero solution.
- 2) $\sigma < m + n + 1$. Then the matrix of the f_{ij} is of rank σ . We may then attribute arbitrary values to $(n + m + 1) - \sigma$ of the η_i and solve (3.4.2) for the remaining η_i .

Lemma 1.2. There exists an $\epsilon > 0$ such that

$$\max_{-a \leq x \leq a} |r_0(x)| < L,$$

where

$$r_0(x) = Q_0(x) - f(x)$$

and $Q_0(x)$ is the rational function which is obtained when the coefficients p_i, q_i , in $Q(x)$ are replaced by

$$p_i - \epsilon \eta_i, q_i - \epsilon \eta_i.$$

Proof. We can divide $[-a, a]$ into two sets of intervals, s_1 and s_2 , where s_1 consists of intervals enclosing the extrema such that $r(x)$ does not vanish in any of these intervals, and s_2 consists of the remaining intervals.

Let us consider a typical interval of s_1 . At the extremum x_j , we have

$$r_0(x_j) = r(x_j) - \epsilon \sum_{i=1}^{m+n+1} \eta_i \frac{\partial r(x_j)}{\partial p_{i-1}} + O(\epsilon^2),$$

(where we are assuming $p_{n+j} = q_j$), since $Q(x)$ - and hence $r(x)$ - is a differentiable function of the parameters p_i so long as the denominator of $Q(x)$ does not vanish in $[-a, a]$. It is easy, but messy, to prove that the term $O(\epsilon^2)$ can be handled. Hence we shall ignore it. By reason of Lemma 1.1 we have

$$r_0(x_j) = r(x_j)[1-\epsilon] + O(\epsilon^2) < r(x_j)$$

since $\epsilon > 0$.

Now let x be any other point of this interval

and let the interval be chosen so that

$$r(x) = r(x_j) + \delta_1$$

which can be written as

$$\sum_i \eta_i \frac{\partial r(x)}{\partial p_i} = \sum_i \eta_i \frac{\partial r(x_j)}{\partial p_i} + \delta_2$$

since $r(x)$ and $\partial r(x)/\partial p_i$ are continuous functions of x .

Now

$$\begin{aligned} r_o(x) &= r(x) - \epsilon \left[\sum_i \eta_i \frac{\partial r(x)}{\partial p_i} \right] + O(\epsilon^2) \\ &= r(x) - \epsilon \left[\sum_i \eta_i \frac{\partial r(x_j)}{\partial p_i} + \delta_2 \right] + O(\epsilon^2), \end{aligned}$$

that is,

$$\begin{aligned} r_o(x) &= r(x) - \epsilon [r(x_j) + \delta_2] + O(\epsilon^2) \\ &= r(x) - \epsilon [r(x) + \delta_1 + \delta_2] + O(\epsilon^2) \\ &= r(x)[1 - \epsilon] - \epsilon(\delta_1 + \delta_2) + O(\epsilon^2) \end{aligned}$$

We require that $|\delta_1 + \delta_2| < 1$ so that we have

$$|r_o(x)| < |r(x)| \text{ and therefore } |r_o(x)| < L.$$

We have yet to fix ϵ . If we let $||r(x)| - L| \leq \delta$ in any interval of s_2 , then it is sufficient to choose ϵ such that $|r_o(x) - r(x)| < \delta$ for all x contained in s_2 .

This completes the proof of Lemma 1.2 and hence the proof of Lemma 1.

We are now in a position to prove Chebyshev's theorem. We will first set up the specific form of equations (3.4.1) when $r(x) = Q(x) - f(x)$ and

$$\begin{aligned} Q(x) &= \frac{p_0 x^n + p_1 x^{n-1} + \dots + p_n}{q_0 x^m + q_1 x^{m-1} + \dots + q_{m-1} x + 1} \\ &= \frac{\sum_{i=0}^n p_i x^{n-i}}{1 + \sum_{i=0}^{m-1} q_i x^{m-i}} \end{aligned}$$

where the abscissae of the extrema of $r(x)$ are x_j , $j = 1, 2, \dots, \sigma$, where $\sigma < m + n + 2$. We have

$$\frac{\partial r(x)}{\partial p_i} = \frac{x^{n-i}}{q(x)} = \frac{q(x)x^{n-i}}{q^2(x)}, \quad i = 0, 1, \dots, n,$$

and

$$\frac{\partial r(x)}{\partial q_i} = - \frac{x^{m-i} p(x)}{q^2(x)}, \quad i = 0, 1, \dots, m-1.$$

Hence the equations (3.4.1) become, in this case,

$$\begin{aligned} \sum_{k=1}^{\sigma} \frac{\lambda_k}{q^2(x_k)} q(x_k) x_k^{n-i} &= 0, \quad i = 0, 1, \dots, n \\ \sum_{k=1}^{\sigma} \frac{\lambda_k}{q^2(x_k)} p(x_k) x_k^{m-i} &= 0, \quad i = 0, 1, \dots, m-1. \end{aligned}$$

Now, if we multiply the first set of equations by β_i , $i = 0, 1, \dots, n$, and the second set of equations by γ_i , $i = 0, 1, \dots, m-1$, and subtract we have

$$(3.4.3) \quad \sum_{k=1}^{\sigma} \frac{\lambda_k}{q^2(x_k)} \Psi(x_k) = 0$$

$$\text{where } \Psi(x) = q(x) \sum_{i=0}^n \beta_{n-i} x^i - p(x) \sum_{i=0}^{m-1} \gamma_{m-1-i} x^{i+1}.$$

It is most convenient to set out the proof as a number of lemmas.

Lemma 2. If $\Psi(x)$ is of order $\sigma - 1$ and if the coefficients β_i , γ_i can be chosen so that

$$(3.4.4) \quad \begin{aligned} \Psi(x) &= (x-x_2)(x-x_3)\dots(x-x_{\sigma}) \\ &= \prod_{j=2}^{\sigma} (x-x_j), \end{aligned}$$

then $\lambda_1 = 0$.

Proof. The lemma is obvious from the form of (3.4.3).

If $\Psi(x)$ is of the form (3.4.4), then $\Psi(x_k)$ vanishes for $k \neq 1$. It follows from (3.4.4) that $\lambda_1 = 0$.

A similar lemma would enable us to show that $\lambda_k = 0$ if $\Psi(x)$ is of the form $\Psi_k(x)$, that is, a product of all the factors $(x-x_j)$, $j = 1, 2, \dots, \sigma$, $j \neq k$.

Lemma 3. $\Psi(x)$ can be reduced to any of the forms $\Psi_k(x)$, $k = 1, 2, \dots, \sigma$ as long as one or more of the coefficients

$$p_0, p_1, \dots, p_{d-1} \quad d = n + m + 2 - \sigma$$

$$q_0, q_1, \dots, q_{d-1}$$

is non-zero.

Proof. We may write

$$(3.4.5) \quad \Psi(x) = q(x) \sum_{i=0}^n \beta_{n-i} x^i - p(x) \sum_{i=0}^{m-1} \gamma_{m-1-i} x^{i+1}.$$

However,

$$\begin{aligned} q(x) \sum_{i=0}^n \beta_{n-i} x^i &= [1 + \sum_{j=0}^{m-1} q_{m-j-1} x^{j+1}] \sum_{i=0}^n \beta_{n-i} x^i \\ &= \sum_{i=0}^n \beta_{n-i} x^i + \sum_{\ell=0}^{m+n-1} \min(\ell, m-1) \sum_{j=0}^{\ell} \beta_{n-\ell+j} q_{m-j-1} x^{\ell+1}, \end{aligned}$$

and

$$\begin{aligned} p(x) \sum_{j=0}^{m-1} \gamma_{m-1-j} x^{j+1} &= [\sum_{i=0}^n p_{n-i} x^i] \sum_{j=0}^{m-1} \gamma_{m-1-j} x^{j+1} \\ &= \sum_{\ell=0}^{m+n-1} \min(\ell, m-1) \sum_{j=0}^{\ell} p_{n-\ell+j} \gamma_{m-1-j} x^{\ell+1} \end{aligned}$$

Hence we can write (3.4.5) as

$$\begin{aligned} \Psi(x) &= \sum_{\ell=0}^{m+n-1} x^{\ell+1} \left[\sum_{j=0}^{\min(\ell, m-1)} \beta_{n-\ell+j} q_{m-j-1} - p_{n-\ell+j} \gamma_{m-1-j} \right] \\ &\quad + \sum_{i=0}^n \beta_{n-i} x^i. \end{aligned}$$

At this stage we find it helpful to write out a few coefficients of this polynomial. Some coefficients are:

$$\begin{aligned} x^{m+n} &: q_0 \beta_0 - p_0 \gamma_0 \\ x^{m+n-1} &: q_1 \beta_0 + q_0 \beta_1 - p_1 \gamma_0 - p_0 \gamma_1 \\ x^{m+n-2} &: q_2 \beta_0 + q_1 \beta_1 + q_0 \beta_2 - p_2 \gamma_0 - p_1 \gamma_1 - p_0 \gamma_2 . \end{aligned}$$

It is now evident that these coefficients form a triangular matrix if we regard the β_i and the γ_i as unknowns and if we equate these coefficients to any chosen polynomial provided that not all the relevant coefficients p_i, q_i are zero.

For example, suppose that $\sigma = m + n + 1$. Then let us try to make $\Psi(x)$ assume the form

$$\begin{aligned} \Psi_1(x) &= (x-x_2)(x-x_3)\dots(x-x_{m+n+1}) \\ &= x^{m+n} - (x_2+\dots+x_{m+n+1}) x^{m+n-1} \dots . \end{aligned}$$

We then have that

$$q_0 \beta_0 - p_0 \gamma_0 = 1$$

$$q_1 \beta_0 + q_0 \beta_1 - p_1 \gamma_0 - p_0 \gamma_1 = -(x_2+x_3+\dots+x_{m+n+1})$$

and we can determine values of β_0, γ_0 to satisfy these

equations unless p_0 and q_0 are both zero. Similarly we may, in this example, make $\Psi(x)$ assume any of the forms $\Psi_k(x)$, $k = 1, \dots, m+n+1$. This would force us to conclude that all the λ_k are zero and by Lemma 1, $Q(x)$ would not be the best approximation. Hence we are forced to conclude that $p_0 = q_0 = 0$ in this example.

We can extend this argument to show that

$$p_0 = p_1 = p_2 = \dots = p_{d-1} = 0 \qquad d = m + n + 2 - \sigma$$

$$q_0 = q_1 = q_2 = \dots = q_{d-1} = 0$$

if there are σ extrema, which completes the proof of Chebyshev's theorem.

It may be noted that the proof of the above theorem does not require that the extrema alternate in sign. It seems that Chebyshev was aware of this property of the rational function of closest approximation; but it is not explicitly stated in his longest paper on this topic (see [2]), and we have been unable to trace an explicit statement of it in the recently published Oeuvres. Priority in the statement of this property is a question of historical interest. It may be noted however that Achiezer's statement of the theorem complements

the above proof since it shows that the extrema of

$$r(x) = Q(x) - f(x)$$

must alternate in sign when $Q(x)$ is the function of closest approximation among the set of functions $p_n(x)/q_m(x)$.

Section 3.5 Rational Approximations to Sets of Ordinates

In this section we shall attack the problem of extending the argument used in the Chebyshev concepts of chapter II, section 6 to the rational approximation case. However, it will be found that the use of determinants is extremely useful, if not necessary. The problem is difficult and only the preliminary work and a start on the solution along with a few examples will be found here.

We will investigate the general approximant

$$(3.5.1) \quad y = \frac{a_n + a_{n-1}x + \dots + a_0x^n}{b_n + b_{n-1}x + \dots + b_0x^n}.$$

For definiteness, we shall base most of our arguments on the very simple approximant

$$(3.5.2) \quad y = \frac{a_1 + a_0x}{b_1 + b_0x},$$

but most of our arguments will apply with little modification to the general approximant. Indeed we shall explicitly point out any limitations. Notice

that no proper rational approximation can be simpler unless it reduces to a polynomial.

We can write (3.5.2) as

$$(3.5.3) \quad a_1 + a_0x - b_1y - b_0xy = 0$$

or as

$$(3.5.4) \quad \begin{bmatrix} 1 & x & y & xy \end{bmatrix} \begin{bmatrix} a_1 \\ a_0 \\ -b_1 \\ -b_0 \end{bmatrix} = 0$$

Notice that (3.5.2) has only three free constants. If the points (x_i, y_i) , $i = 1, 2, 3$, lie on (3.5.2), then we have from (3.5.4)

$$(3.5.5) \quad \begin{bmatrix} 1 & x_1 & y_1 & x_1y_1 \\ 1 & x_2 & y_2 & x_2y_2 \\ 1 & x_3 & y_3 & x_3y_3 \\ 1 & x & y & xy \end{bmatrix} \begin{bmatrix} a_1 \\ a_0 \\ -b_1 \\ -b_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

the last equation of (3.5.5) expressing the fact that (3.5.4) is true for any (x, y) on the curve.

The determinant of (3.5.5) must vanish. Hence

the equation of the curve of form (3.5.2) through (x_i, y_i) , $i = 1, 2, 3$, is

$$(3.5.6) \quad \begin{vmatrix} 1 & x_1 & y_1 & x_1 y_1 \\ 1 & x_2 & y_2 & x_2 y_2 \\ 1 & x_3 & y_3 & x_3 y_3 \\ 1 & x & y & x y \end{vmatrix} = 0$$

Clearly, our reasoning extends to higher order approximants.

Now our aim is to approximate an arbitrary continuous function by a suitable rational function; but this problem is much too difficult to encompass in a single step. We begin with a much simpler problem. To approximate $2n + 2$ ordinates by a rational function of the type

$$(3.5.7) \quad y = \frac{a_n + a_{n-1}x + \dots + a_0 x^n}{b_n + b_{n-1}x + \dots + b_0 x^n}$$

Since the degree of the polynomial equation for p is one greater than the degree of the denominator, the odd degree denominator yields a more difficult problem in some respects. Hence we may as well face at once

the difficulty of an odd degree denominator.

In keeping with our purpose of simplicity of presentation, we shall base most of our arguments on the simple approximant (3.5.2), which corresponds to $n = 1$.

Suppose we use (3.5.5) to approximate four ordinates y_i , $i = 1, 2, 3, 4$. Obviously we cannot fit them exactly by a curve of the form (3.5.2). Therefore let us use the notation

$$y_i + \rho f_i(v_i), \quad i = 1, 2, 3, 4,$$

for the ordinates of the approximant.

$$y = \frac{a_1 + a_0 x}{b_1 + b_0 x}$$

at the points x_i , $i = 1, 2, 3, 4$. The functions $f_i(v_i)$ are continuous differentiable functions of their arguments and vary continuously between ± 1 . We wish to emphasize by the notation that the results we shall obtain are quite general and in no way restricted to any choice of function.

Since the maximum deviation between the ordinates and the approximate ordinates is $\pm \rho_1$, at least one of the $f_i(v_i)$ must be ± 1 .

From (3.5.6), the ordinates $y_1 + f_1(v_1)$ are subject to the equation

$$(3.5.8) \quad F(\rho, v_1) = \begin{vmatrix} 1 & x_1 & y_1 + f_1(v_1)\rho & x_1(y_1 + \rho f_1(v_1)) \\ 1 & x_2 & y_2 + f_2(v_2)\rho & x_2(y_2 + \rho f_2(v_2)) \\ 1 & x_3 & y_3 + f_3(v_3)\rho & x_3(y_3 + \rho f_3(v_3)) \\ 1 & x_4 & y_4 + f_4(v_4)\rho & x_4(y_4 + \rho f_4(v_4)) \end{vmatrix} = 0$$

We now seek to minimize ρ with respect to the v_1 . The necessary conditions are of the form

$$(3.5.9) \quad \begin{vmatrix} 1 & x_1 & \rho f_1'(v_1) & x_1(y_1 + \rho f_1(v_1)) \\ 1 & x_2 & 0 & x_2(y_2 + \rho f_2(v_2)) \\ 1 & x_3 & 0 & x_3(y_3 + \rho f_3(v_3)) \\ 1 & x_4 & 0 & x_4(y_4 + \rho f_4(v_4)) \end{vmatrix} + \begin{vmatrix} 1 & x_1 & y_1 + \rho f_1(v_1) & \rho x_1 f_1'(v_1) \\ 1 & x_2 & y_2 + \rho f_2(v_2) & 0 \\ 1 & x_3 & y_3 + \rho f_3(v_3) & 0 \\ 1 & x_4 & y_4 + \rho f_4(v_4) & 0 \end{vmatrix} = 0$$

where we have differentiated with respect to v_1 . Three similar equations can be obtained by differentiating with respect to v_2 , v_3 and v_4 . At a maximum we must have

$$\frac{\partial \rho}{\partial v_i} = 0$$

but since $F_\rho d\rho + F_{v_1} dv_1 = 0$,

$$\frac{\partial \rho}{\partial v_1} = - \frac{F_{v_1}}{F_\rho} ;$$

that is, we must have $F_{v_1} = 0$, $F_\rho \neq 0$. Now

(3.5.9) is satisfied by either of the following

- (a) $f_1'(v_1) = 0$ that is, at the maximum or minimum value of $f_1(v_1)$
- (b) $f_1'(v_1) \neq 0$ but the sum of the determinants is zero;

and similarly for the other three equations of this type.

Let us take hypothesis (b) first. If we multiply the equations of type (3.5.9) by $f_i(v_i)/\rho f_i'(v_i)$ and add the four we obtain

(3.5.10)

$$\begin{vmatrix} 1 & x_1 & f_1(v_1) & x_1(y_1 + \rho f_1(v_1)) \\ 1 & x_2 & f_2(v_2) & x_2(y_2 + \rho f_2(v_2)) \\ 1 & x_3 & f_3(v_3) & x_3(y_3 + \rho f_3(v_3)) \\ 1 & x_4 & f_4(v_4) & x_4(y_4 + \rho f_4(v_4)) \end{vmatrix} + \begin{vmatrix} 1 & x_1 & y_1 + \rho f_1(v_1) & x_1 f_1(v_1) \\ 1 & x_2 & y_2 + \rho f_2(v_2) & x_2 f_2(v_2) \\ 1 & x_3 & y_3 + \rho f_3(v_3) & x_3 f_3(v_3) \\ 1 & x_4 & y_4 + \rho f_4(v_4) & x_4 f_4(v_4) \end{vmatrix} = 0$$

The L. H. S. of (3.5.10) is simply F_ρ . Hence hypothesis (b) does not yield the necessary conditions for a minimum.

Thus our necessary conditions for a minimum must be obtained from hypothesis (a); that is

$$f'_1(v_1) = 0 \quad \text{or} \quad f_1(v_1) = \pm 1.$$

We can now write (3.5.8) in the form

(3.5.11)

$$\begin{vmatrix} 1 & x_1 & y_1 + \epsilon_1 \rho & x_1(y_1 + \epsilon_1 \rho) \\ 1 & x_2 & y_2 + \epsilon_2 \rho & x_2(y_2 + \epsilon_2 \rho) \\ 1 & x_3 & y_3 + \epsilon_3 \rho & x_3(y_3 + \epsilon_3 \rho) \\ 1 & x_4 & y_4 + \epsilon_4 \rho & x_4(y_4 + \epsilon_4 \rho) \end{vmatrix} = 0$$

where the ϵ_i are ± 1 - we don't know which.

It might appear at this point that we ought to demonstrate the existence of a minimum. However we shall defer this for the present since we have no knowledge of the kind of difficulty which an existence proof will encounter and must surmount. Actually, there is a perfectly good existence proof which depends on continuity considerations but we should like to achieve a proof which rests on simpler considerations.

Let us now turn to the question of uniqueness. Suppose we have a function of the type (3.5.2) which

approximates the ordinates y_i , $i = 1, 2, 3, 4$, by the ordinates $y_1 + \rho$, $y_2 - \rho$, $y_3 + \rho$, $y_4 - \rho$, that is suppose the deviations at the points (x_i, y_i) are of equal magnitude and of alternating sign.

Suppose also that the denominator $b_1 + b_0 x$ does not vanish in the interval x_1 to x_4 - we assume that the x_i are arranged in ascending order of magnitude.

Then we assert there can be only one function of the above type. For suppose there are two different solutions of (3.5.11), say $p(x)/q(x)$ and $\bar{p}(x)/\bar{q}(x)$ with maximum deviation ρ , $\bar{\rho}$ respectively. Now ρ , $\bar{\rho}$ must be different for otherwise the rational functions would be equal, too. For x_i , $i = 1, 2, 3, 4$, the difference

$$\frac{p(x)}{q(x)} - \frac{\bar{p}(x)}{\bar{q}(x)} = \frac{p\bar{q} - \bar{p}q}{q\bar{q}}$$

assumes the values $(-1)^i(\rho - \bar{\rho}) \neq 0$, thus being subjected to at least three changes of sign. The numerator is of degree two, and can account for at most two sign changes. Hence this leaves at least one sign change to the denominator; that is, one of the approximations must have a pole within the range of approximation.

The argument can easily be generalized.

We have no reason as yet to favour the existence of one approximant of a given type nor the existence of another approximant of this type. It is sufficient that coexistence is denied. Hence, if we can demonstrate the existence of any one approximant, the argument is complete. In particular, if an approximation of equal and opposite deviation exists, we can do no better.

To demonstrate existence algebraically, we shall have to examine closely determinants of the type (3.5.11).

At this point it appears helpful to work out a few examples to fix the above concepts.

Section 3.5.2 Examples

Example 1 To fit an approximation of the type

$$y = \frac{a_1 + a_0 x}{b_1 + b_0 x}$$

to the points

| | | | | |
|---|----|---|---|---|
| x | 0 | 1 | 2 | 3 |
| y | -1 | 0 | 2 | 1 |

We will look for an approximation of equal and alternating deviations, $\pm \rho$. Employing (3.5.11) with $\epsilon_1 = (-1)^{1-1}$ and evaluating for ρ , we obtain $\rho = \pm \sqrt{7}/4$. This would contradict our general argument if both values of ρ were admissible. Let us first take $\rho = -\sqrt{7}/4$.

The curve through the points

| x | 0 | 1 | 2 |
|---|-------------------|--------------|------------------|
| y | $-1 - \sqrt{7}/4$ | $\sqrt{7}/4$ | $2 - \sqrt{7}/4$ |

is

$$\begin{vmatrix} 1 & 0 & -1 - \sqrt{7}/4 & 0 \\ 1 & 1 & +\sqrt{7}/4 & \sqrt{7}/4 \\ 1 & 2 & 2 - \sqrt{7}/4 & 4 - \sqrt{7}/2 \\ 1 & x & y & xy \end{vmatrix} = 0$$

that is

$$y = \frac{3}{4} \frac{(3 + \sqrt{7})x - 3}{(4 - \sqrt{7}) + (\sqrt{7} - 1)x}$$

The crucial point is the zero of the denominator at

$$x = - \frac{(4 - \sqrt{7})}{(\sqrt{7} - 1)}$$

which lies outside the interval $[0, 3]$. Hence the approximation is acceptable.

Let us investigate the value $\rho = +\sqrt{7}/4$. The curve through the points

| | | | |
|---|-----------------|---------------|----------------|
| x | 0 | 1 | 2 |
| y | $-1+\sqrt{7}/4$ | $-\sqrt{7}/4$ | $2+\sqrt{7}/4$ |

is

$$\begin{vmatrix} 1 & 0 & -1+\sqrt{7}/4 & 0 \\ 1 & 1 & -\sqrt{7}/4 & -\sqrt{7}/4 \\ 1 & 2 & 2+\sqrt{7}/4 & 4+\sqrt{7}/2 \\ 1 & x & y & xy \end{vmatrix} = 0.$$

If we evaluate the determinant we find

$$y = \frac{(\frac{9}{4} - \frac{3\sqrt{7}}{4})x - \frac{9}{4}}{4 + \sqrt{7} - (1+\sqrt{7})x}.$$

The denominator has a zero at

$$x = \frac{4 + \sqrt{7}}{1 + \sqrt{7}} \simeq 1.8$$

That is, the approximation becomes infinite within the interval $(0,3)$ which is in agreement with our uniqueness argument, since only one of the approximations fulfills the stated conditions.

In this simple case we have demonstrated the

existence of an approximation of the prescribed type with the necessary properties. From the denial of co-existence, we can do no better.

We are now fairly well alerted to the difficulties which any proofs along this line must surmount. However, let us look at another example which exhibits the difficulty in a different way.

Example 2 To fit the four points

| | | | | |
|---|----|---|---|---|
| x | 0 | 1 | 2 | 3 |
| y | -1 | 0 | 2 | 5 |

approximately by

$$y = \frac{ax + b}{cx + d}.$$

Using (3.5.11) with $\epsilon_i = (-1)^{i-1}$, we obtain

$$\rho = +1 \pm \frac{\sqrt{19}}{4} \simeq +2.09, -0.09$$

A plausible guess is that the smaller absolute value of ρ is the one that we want, that is

$$\rho = +1 - \frac{\sqrt{19}}{4}$$

The curve corresponding to this value of ρ is

$$\begin{vmatrix} 1 & 0 & -\sqrt{19}/4 & 0 \\ 1 & 1 & -1+\sqrt{19}/4 & \sqrt{19}/4 - 1 \\ 1 & 2 & 3-\sqrt{19}/4 & 6-\sqrt{19}/2 \\ 1 & x & y & xy \end{vmatrix} = 0$$

Thus

$$y = \frac{1}{4} \frac{-(8\sqrt{19} - 19) + (11\sqrt{19} - 31)x}{(8 - \sqrt{19}) - (5 - \sqrt{19})x},$$

where the pole is at $x = \frac{7 + \sqrt{19}}{2} \simeq \frac{11.36}{2}$, which is outside the range $[0, 3]$.

Let us now test the other value of ρ , where $\rho = +1 + \frac{\sqrt{19}}{4}$. The curve corresponding to this value of ρ is

$$\begin{vmatrix} 1 & 0 & \sqrt{19}/4 & 0 \\ 1 & 1 & -(1+\sqrt{19}/4) & -(1+\sqrt{19}/4) \\ 1 & 2 & 3+\sqrt{19}/4 & 6+\sqrt{19}/2 \\ 1 & x & y & xy \end{vmatrix} = 0.$$

Thus

$$y = \frac{1}{4} \frac{(8\sqrt{19} + 19) - (11\sqrt{19} + 31)x}{(8 + \sqrt{19}) - (5 + \sqrt{19})x}$$

where the pole is at $x = \frac{7 - \sqrt{19}}{2} \simeq 1.32$, which is within the range $[0,3]$. Therefore this approximation is not acceptable. Let us take one more example.

Example 3 To fit the four points

| | | | | |
|---|----|---|----|---|
| x | 0 | 1 | 2 | 3 |
| y | -1 | 2 | -2 | 5 |

by $y = \frac{ax + b}{cx + d}$ using (3.5.11) with $\epsilon_i = (-1)^{i-1}$,

we have $\rho = +\frac{5}{2} \pm \frac{\sqrt{13}}{4} \simeq +1.6, +3.4$. Taking the least absolute value of ρ first we have the curve given by:

$$\begin{vmatrix} 1 & 0 & \frac{3}{2} - \frac{1}{4}\sqrt{13} & 0 \\ 1 & 1 & -\frac{1}{2} + \frac{1}{4}\sqrt{13} & -\frac{1}{2} + \frac{1}{4}\sqrt{13} \\ 1 & 2 & \frac{1}{2} - \frac{1}{4}\sqrt{13} & 1 - \frac{1}{2}\sqrt{13} \\ 1 & x & y & xy \end{vmatrix} = 0$$

Thus

$$y = \frac{1}{4} \frac{-8\sqrt{13} + 25 + (7\sqrt{13} - 23)x}{2 - \sqrt{13} - (3 - \sqrt{13})x}$$

where the pole is at $x = \frac{7 + \sqrt{13}}{4} \simeq 2.65$ which is within the range $[0,3]$.

Now let us examine the other value of ρ ,
where the curve is given by:

$$\begin{vmatrix} 1 & 0 & \frac{3}{2} + \frac{1}{4}\sqrt{13} & 0 \\ 1 & 1 & -\frac{1}{2} - \frac{1}{4}\sqrt{13} & -\frac{1}{2} - \frac{1}{4}\sqrt{13} \\ 1 & 2 & \frac{1}{2} + \frac{1}{4}\sqrt{13} & 1 + \frac{1}{2}\sqrt{13} \\ 1 & x & y & xy \end{vmatrix} = 0.$$

Thus

$$y = \frac{1}{4} \frac{(8\sqrt{13} + 25) - (7\sqrt{13} + 23)x}{\sqrt{13} + 2 - (\sqrt{13} + 3)x}$$

where the pole is at $x = \frac{7 - \sqrt{13}}{4} \simeq 0.85$ which
lies within the range $[0, 3]$.

Thus both of our approximations in this example
have poles within the interval of approximation. Hence
we have been unable to find an acceptable rational
approximation of the required form to the given ordinates.

We refuse to accept approximation with poles,
because we are trying to approximate - in the end -
finite continuous functions.

Section 3.5.3 Interpretation of the Examples

In examples 1 and 2 there was only one admissible

approximation which is comforting.

In example 3 there was no admissible approximation; both approximations had poles within the strip of points. This example is of considerable theoretical importance since it denies the assertion that an admissible approximation of the specified form to a specific set of points always exists.

We can now see what ingredients must enter any satisfactory proof along algebraic lines: we must be able to find a real root of the equation for ρ , say $F(\rho) = 0$, which is not a root of the denominator, say $G(\rho, x)$, for $x_1 \leq x \leq x_n$. One way of doing this would be to show that there exists real values of x such that $F(\rho) = (a\rho + b) G(\rho, x)$, where $x_1 \leq x \leq x_n$ and a, b are constants. However, this does not appear to be easy.

Example 3 implies one other important point: if the ordinates y_1, y_2, y_3, y_4 are such that, by adding a suitable constant, γ , they can be made to alternate in sign, there is no real root of $F(\rho) = 0$ unless ρ is sufficiently large to change at least one alternation in sign. Hence, if we suspect the existence of a real ρ , we can bound it by

$$\min_i |y_i + \gamma| < \rho < \max_i |y_i + \gamma| .$$

A possible generalization of this could be: if $F(\rho)$ is an even-order polynomial in ρ and if the ordinates, $y_1, y_2, \dots, y_{2n+2}$ are such that, by adding a suitable constant, γ , they can be made to alternate in sign, there is no real root of $F(\rho) = 0$ unless ρ is sufficiently large to change at least one alternation in sign.

Our examples do not imply any contradiction of the existence theorem for rational approximation. Rather, they imply that certain sets of abscissae in the interval of approximation may be excluded in the search for extrema. Werner has studied rational approximations of the form (3.5.2) to a continuous function $f(x)$. He gives the following theorem (see [24]):

Theorem 3.5.3.1 Let (x_i, f_i) , $i = 0, 1, 2, 3$, be given and let $x_{i+1} > x_i$. The necessary and sufficient condition for the existence in $[x_0, x_3]$ of a continuous rational approximation such that

$$Q(x_i) - f_i = (-1)^i [Q(x_0) - f_0]$$

is

$$\operatorname{sgn}(f_0 - f_2) = \operatorname{sgn}(f_1 - f_3) .$$

It appears that this theorem can be generalized to include higher-order rational approximations, although this has yet to be done.

We will now consider two methods for obtaining ρ . It will be found that the second method gives ρ as the eigenvalues of a symmetrical matrix with the associated eigenvectors giving the coefficients of the denominator corresponding to that value of ρ .

Section 3.5.4 Methods for Obtaining ρ

Let us write the determinant (3.5.8) - in generalized form - as

(3.5.12)

$$| 1, x_i, x_i^2, \dots, x_i^n, \eta_i, x_i \eta_i, \dots, x_i^m \eta_i | \quad ,$$

$$i = 1, \dots, n+m+2$$

where $\eta_i = y_i + (-1)^{i-1} \rho$. It is helpful to work with matrices rather than determinants here. We shall use the notation

X_n for the submatrix $[1, x, \dots, x^n]$ of order $(m+n+2) \times (n+1)$,

X_m for the submatrix $[1, x, \dots, x^m]$ of order $(m+n+2) \times (m+1)$,

Y for the diagonal matrix of order $(m+n+2) \times (m+n+2)$ whose diagonal elements are y_i , and

D for the diagonal matrix of order $(m+n+2) \times (m+n+2)$ whose diagonal elements are $(-1)^{i-1}$.

The above determinant (3.5.12) is then the determinant of the matrix $[X_n, (Y + \rho D)X_m]$ and can be written as $|X_n, (Y + \rho D)X_m|$. Now we use the equation

$$|X_n, (Y + \rho D)X_m| = 0$$

as an equation of determine ρ . Clearly it is equivalent to a polynomial equation of order $m+1$. Hence it would be helpful to reduce the determinant to an equivalent determinant of order $m+1$. Two methods of reduction are given below.

First Method

It can easily be seen that the set of $m+n+2$ vectors X_n, DX_m are linearly independent. Let us find the inverse of $[X_n \mid DX_m]$ partitioning the matrix by columns; the inverse will be conformably partitioned by rows and can be denoted by

$$\begin{bmatrix} -\frac{\alpha}{\beta} \end{bmatrix}.$$

Hence we have

$$\begin{bmatrix} -\frac{\alpha}{\beta} \\ 0 \end{bmatrix} \begin{bmatrix} X_n \\ DX_m \end{bmatrix} = \begin{bmatrix} -\frac{I}{0} & -\frac{0}{I} \end{bmatrix},$$

that is

$$\begin{aligned} \alpha X_n &= I & \alpha DX_m &= 0 \\ \beta X_n &= 0 & \beta DX_m &= I. \end{aligned}$$

Hence if we premultiply the matrix

$$\begin{bmatrix} X_n \\ (Y + \rho D)X_m \end{bmatrix} \quad \text{by}$$

$$\begin{bmatrix} -\frac{\alpha}{\beta} \\ 0 \end{bmatrix}$$

we obtain

$$\begin{bmatrix} -\frac{I}{0} & -\frac{\alpha Y X_m}{\beta Y X_m + \rho I} \end{bmatrix}$$

and the determinantal equation

$$| X_n, (Y + \rho D)X_m | = 0$$

is equivalent to

$$| \beta Y X_m + \rho I | = 0$$

which is of order $m+1$.

This method is easy to understand and simple to use, however it is likely to be ill-conditioned for large n or m , since Vandermonde determinants are ill-conditioned and the determinants of the matrices X_m and X_n are Vandermonde determinants.

Second Method

We will find it convenient to prove some lemmas first.

Lemma 3 The $(p,q)^{th}$ element of the matrix

$X_m^T \Lambda X_n - \Lambda$ is a diagonal matrix of order $(m+n+2) \times (m+n+2)$ whose diagonal elements are λ_i - is $\sum_{i=1}^{m+n+2} \lambda_i x_i^{p+q-2}$.

Proof: We have $X_m = [1 \ x \ x^2 \ \dots \ x^m]$. Hence

$$X_m^T = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^m \end{bmatrix}$$

and since

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \lambda_2 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & & & \cdot & \\ 0 & 0 & & & & \lambda_{m+n+2} \end{bmatrix}$$

the p^{th} row of $X_m^T \Lambda$ has elements equal to $x_i^{p-1} \lambda_i$.

Now $X_n = [1 \ x \ x^2 \ . \ . \ . \ x^n]$ and hence the (p,q) element of $X_m^T \Lambda X_n$ is

$$x_1^{p-1} \lambda_1 x_1^{q-1} + x_2^{p-1} \lambda_2 x_2^{q-1} + \dots = \sum_{i=1}^{m+n+2} \lambda_i x_i^{p+q-2}.$$

Lemma 4 The elements of diagonal matrix Λ can be chosen such that $X_m^T \Lambda X_n = 0$.

Proof: The lemma is equivalent to the statement that λ_i , $i = 1, 2, \dots, m+n+2$, can be found such that

$$\sum_{i=1}^{m+n+2} \lambda_i x_i^s = 0, \quad s = 0, 1, \dots, m+n.$$

Now it is well known that any $(\ell+1)$ vectors in ℓ -space are linearly dependant. Hence the $(\ell+1)$ vectors x_i^{s-1} , $s = 1, 2, \dots, \ell$ are linearly dependant if $x_i \neq x_j$. We shall suppose that $x_{i+1} > x_i$. The equations are

$$\begin{bmatrix} 1 & . & . & . & 1 \\ x_1 & . & . & . & x_{\ell+1} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ x_1^{\ell-1} & & & & x_{\ell+1}^{\ell-1} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ . \\ . \\ . \\ \lambda_{\ell+1} \end{bmatrix} = 0$$

and can be written as

$$\begin{bmatrix} 1 & . & . & . & 1 \\ x_2 & & & & x_{\ell+1} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ x_2^{\ell-1} & & & & x_{\ell+1}^{\ell-1} \end{bmatrix} \begin{bmatrix} \lambda_2 \\ \lambda_3 \\ . \\ . \\ . \\ \lambda_{\ell+1} \end{bmatrix} = -\lambda_1 \begin{bmatrix} 1 \\ x_1 \\ . \\ . \\ . \\ x_1^{\ell-1} \end{bmatrix}$$

which we can write as

$$\begin{bmatrix} 1 & . & . & . & 1 \\ x_2 & & & & x_{\ell+1} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ x_2^{\ell-1} & & & & x_{\ell+1}^{\ell-1} \end{bmatrix} \begin{bmatrix} -\lambda_2/\lambda_1 \\ -\lambda_3/\lambda_1 \\ . \\ . \\ . \\ -\lambda_{\ell+1}/\lambda_1 \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ . \\ . \\ . \\ x_1^{\ell-1} \end{bmatrix}$$

Using Cramer's rule, we can write

$$-\frac{\lambda_k}{\lambda_1} = \frac{\begin{vmatrix} 1 & . & . & . & 1 & 1 & 1 & . & . & . & 1 \\ x_2 & & & & x_{k-1} & x_1 & x_{k+1} & & & & x_{\ell+1} \\ . & & & & . & . & . & & & & . \\ . & & & & . & . & . & & & & . \\ . & & & & . & . & . & & & & . \\ x_2^{\ell-1} & & & & x_{k-1}^{\ell-1} & x_1^{\ell-1} & x_{k+1}^{\ell-1} & & & & x_{\ell+1}^{\ell-1} \end{vmatrix}}{\begin{vmatrix} 1 & 1 & . & . & . & 1 \\ x_2 & x_3 & & & & x_{\ell+1} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ x_2^{\ell-1} & x_3^{\ell-1} & & & & x_{\ell+1}^{\ell-1} \end{vmatrix}}$$

Now the above determinants are of the form of the Vandermonde determinants which can be written as

$$\begin{vmatrix} 1 & 1 & . & . & . & 1 \\ x_1 & x_2 & & & & x_{n+1} \\ x_1^2 & x_2^2 & & & & x_{n+1}^2 \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ x_1^n & x_2^n & & & & x_{n+1}^n \end{vmatrix}$$

and whose value is $\prod_{\substack{j=2 \\ i=1 \\ i < j}}^{n+1} (x_j - x_i)$. Hence we have

$$-\lambda_k / \lambda_1 = (-1)^{k-2} \frac{\prod_{\substack{i=2 \\ j=1 \\ j < i \\ j, i \neq k}}^{\ell+1} (x_i - x_j)}{\prod_{\substack{i=3 \\ j=2 \\ j < i}}^{\ell+1} (x_i - x_j)}$$

$$= \frac{\prod_{\substack{i=2 \\ i \neq k}}^{\ell+1} (x_i - x_1)}{\prod_{\substack{i=2 \\ i \neq k}}^{\ell+1} (x_i - x_k)}$$

$$= (-1) \frac{\prod_{i=2}^{\ell+1} (x_i - x_1)}{\prod_{\substack{i=1 \\ i \neq k}}^{\ell+1} (x_i - x_k)}.$$

Therefore $\lambda_k = \prod_{\substack{i=1 \\ i \neq k}}^{\ell+1} (x_i - x_k)^{-1}$, $k = 1, 2, \dots, \ell+1$.

Now since $x_{i+1} > x_i$ then if

λ_1 is positive we have

λ_2 negative,

λ_3 positive,

and so on.

Hence for $x_{i+1} > x_i$, $\text{sign}(\lambda_{i+1}) = -\text{sign}(\lambda_i)$.

This completes the proof of Lemma 4.

If the matrix $[X_n, (Y + \rho D)X_m]$ is premultiplied by $X_m^T \Lambda$, where the λ_i are chosen in accordance with Lemma 4, we have

$$\begin{aligned} (3.5.13) \quad X_m^T \Lambda [X_n, (Y + \rho D)X_m] \\ = [0, X_m^T \Lambda Y X_m + \rho X_m^T \Lambda D X_m] \end{aligned}$$

and the determinantal equation

$$|X_n, (Y + \rho D)X_m| = 0$$

is equivalent to

$$(3.5.14) \quad |X_m^T \Lambda Y X_m + \rho X_m^T \Lambda D X_m| = 0.$$

Since $(D)_{ij} = (-1)^{i-1} \delta_{ij}$, where δ_{ij} is the Kronecker delta, we have from Lemma 4 that the diagonal elements of ΛD are positive. Hence the second of the above matrices

is symmetrical and positive definite while the first is clearly symmetrical. Thus we can write

$$X_m^T \Lambda D X_m = L L^T$$

where L is real. Premultiplying (3.5.14) by L^{-1} and post-multiplying by $(L^T)^{-1}$, we have

$$(3.5.15) \quad | L^{-1} X_m^T \Lambda Y X_m (L^T)^{-1} + \rho I | = 0.$$

All the eigenvalues of the above matrix are real since it is symmetric. Also, its eigenvectors are orthogonal.

Denoting the eigenvectors of the matrix corresponding to the determinant (3.5.15) by r_j , $j = 1, \dots, m+1$, the eigenvectors of the matrix of (3.5.14) are given by

$$r_j = L^T q_j$$

and the orthogonality relation $r_k^T r_j = 0$ is equivalent to $q_k^T L L^T q_j = 0$ for $j \neq k$. But,

$$(3.5.16) \quad \begin{aligned} q_k^T L L^T q_j &= q_k^T X_m^T \Lambda D X_m q_j \\ &= \sum_{i=1}^{m+n+2} |\lambda_i| q_k(x_i) q_j(x_i) = 0 \end{aligned}$$

where $q_k(x) = a_{m,k} + a_{m-1,k}x + \dots + a_{0,k}x^m$.

Now, from (3.5.16) it is impossible that $q_k(x)$ and $q_j(x)$ should both be one-signed at each of the points x_i , $i = 1, 2, \dots, m+n+2$; for if so, the L.H.S. of (3.5.16) would be positive or negative. Hence, only one, at most, of the q_k vectors can correspond to a denominator of $Q(x)$ which is one-signed throughout the interval of approximation.

CHAPTER IV

METHODS OF FINDING RATIONAL INTERPOLATION FORMULAE

Section 4.1 Introduction

Although this thesis is mainly concerned with "best-fit" - in the Chebyshev sense - rational approximations, this chapter is devoted to a few of the algorithms used to find rational interpolation formulae, since there is still a definite need for interpolatory rational approximations. The algorithms will be preceded by a brief discussion of the relationship between continued fractions and rational functions. The algorithms given below are those that give interpolatory rational approximations in either continued-fraction form, or proper rational function form - that is, a rational function with a polynomial denominator.

Section 4.2 Some Relationships between Finite Continued Fractions and Rational Functions

The form of a typical continued fraction is

$$(4.2.1) \quad y = a_0 + \frac{b_1}{a_1 + \frac{b_2}{a_2 + \frac{b_3}{a_3 + \dots}}}$$

which is commonly written as:

$$(4.2.2) \quad y = a_0 + \frac{b_1}{a_1} + \frac{b_2}{a_2} + \frac{b_3}{a_3} + \dots$$

A convenient method of examining (4.2.1) or (4.2.2) is to express the continued fraction as a rational function, where the successive higher-order rational functions are called convergents. The first few convergents are:

$$\frac{p_0}{q_0} = \frac{a_0}{1} \qquad \frac{p_1}{q_1} = \frac{a_0 a_1 + b_1}{a_1}$$

$$\frac{p_2}{q_2} = \frac{a_0 a_1 a_2 + a_0 b_2 + a_2 b_1}{a_1 a_2 + b_2}$$

Examining the mode of formation of these rational expressions we see that successive convergents can be obtained from the difference equations

$$(4.2.3) \quad \begin{aligned} p_n &= a_n p_{n-1} + b_n p_{n-2} \\ q_n &= a_n q_{n-1} + b_n q_{n-2} \end{aligned}$$

If we replace the b_i by $c_i(x-x_i)$ we see that for odd n the n^{th} convergent has the same power of x in the numerator as it has in the denominator, and for even n the degree of the numerator is one greater than the degree of the denominator. Hence we can write

$$\begin{aligned}
 (4.2.4) \quad r_n(x) &= \frac{\alpha_0 + \alpha_1 x + \dots + \alpha_p x^p}{\beta_0 + \beta_1 x + \dots + \beta_{p-1} x^{p-1}}, \quad n = 2p \\
 r_n(x) &= \frac{\alpha_0 + \alpha_1 x + \dots + \alpha_p x^p}{\beta_0 + \beta_1 x + \dots + \beta_p x^p}, \quad n = 2p+1
 \end{aligned}$$

Since the numerator and denominator of either form of (4.2.4) can be divided through by any one of the non-zero coefficients, the first form involves $2p$ independent parameters and the second form $2p+1$ such parameters, so that in either case n independent constants are available for the determination of the approximation. Hence the n^{th} convergent gives an approximation to $f(x)$ by a rational function of x which, in general, agrees with $f(x)$ at the n points x_0, x_1, \dots, x_{n-1} if $a_{n-1} \neq 0$ and if all preceding a 's are finite.

The uniqueness of a rational function of a given form passing through the points (x_i, y_i) , $i = 1, 2, \dots, n$, is easily shown. We proceed thus: Let us denote the rational function passing through the given points by

$$\frac{p_n(x)}{q_n(x)}.$$

Suppose that there is another rational function:

$$\frac{\bar{p}_n(x)}{\bar{q}_n(x)}$$

such that

$$\frac{\bar{p}_n(x_i)}{\bar{q}_n(x_i)} = y_i, \quad i = 1, 2, \dots, n.$$

Then $\psi(x) = p_n(x_i) \bar{q}_n(x_i) - \bar{p}_n(x_i) q_n(x_i) = 0$,
 $i = 1, 2, \dots, n$. (We are assuming here that
 $q_n(x_i) \neq 0$, $\bar{q}_n(x_i) \neq 0$.) That is, the polynomial
 $\psi(x)$ vanishes for n values of x . Whether n is even or
 odd, the degree of $\psi(x)$ is $n-1$. Hence $\psi(x)$ must be
 identically zero. Therefore $p_n(x)/q_n(x)$ and
 $\bar{p}_n(x)/\bar{q}_n(x)$ are identical except perhaps for a factor;
 that is $\bar{p}_n(x)/\bar{q}_n(x)$ may be of the form:

$$\frac{A(x)p_n(x)}{A(x)q_n(x)}.$$

However, we may find that the rational
 function - of a specified form - which is to pass
 through the n points, has a pole within the interval of
 interest, or leads to an inconsistent set of equations

for the determination of the coefficients of the function.
 For example finding the rational function that passes
 through $(-1,-1)$ and $(1,1)$ and has the form

$$\frac{a}{x + b}$$

we obtain $a = 1$, $b = 0$ and the function has a pole at
 $x = 0$. If we take the function to be of the form

$$\frac{a}{bx + 1}$$

we obtain the equations

$$a - b = -1$$

$$a - b = 1$$

which are inconsistent.

Section 4.3 The Inverse Divided - Difference Interpolation

Formula

We can consider Newton's divided-difference
 polynomial interpolation formula (Chapter 2, section 4),
 with an error term, as the identity which results from
 writing

$$(4.3.1) \quad f(x) = u_0(x)$$

and performing the successive substitutions

$$(4.3.2) \quad u_k(x) = u_k(x_k) + (x-x_k)u_{k+1}(x),$$

$$k = 0, 1, \dots, n-1,$$

with the abbreviation

$$(4.3.3) \quad u_k(x) = f[x_0, x_1, \dots, x_{k-1}, x].$$

The algorithm for the calculation of the successive divided differences follows directly from (4.2.2) and (4.2.3), with $x = x_k$, in the form

$$(4.3.4) \quad f[x_0, x_1, \dots, x_{k-2}, x_{k-1}, x_k] =$$

$$\frac{f[x_0, x_1, \dots, x_{k-2}, x_k] - f[x_0, \dots, x_{k-2}, x_{k-1}]}{x_k - x_{k-1}}$$

The result of assuming that the $(n+1)^{\text{th}}$ divided difference $u_{n+1}(x)$ is identically zero (or that the n^{th} divided difference is constant) is the equation of the polynomial $y(x)$, of degree $\leq n$, which agrees with $f(x)$ at the $n+1$ points x_0, x_1, \dots, x_n . If $u_{n+1}(x)$ actually vanishes identically, then $y(x)$ is identically equal to $f(x)$.

A variety of other identities may be obtained in a similar way by making use of other sets of transformations. In particular, the substitution sequence

$$f(x) = v_0(x)$$

$$(4.3.5) \quad v_k(x) = v_k(x_k) + \frac{x-x_k}{v_{k+1}(x)}, \quad k = 0, 1, 2, \dots$$

leads to the following result

$$(4.3.6) \quad f(x) = v_0(x_0) + \frac{x-x_0}{v_1(x_1)} + \frac{x-x_1}{v_2(x_2)} + \frac{x-x_2}{v_3(x_3)}$$

if truncated after three terms. Thus, more generally, we are led to the continued-fraction representation

$$(4.3.7) \quad f(x) = a_0 + \frac{x-x_0}{a_1} + \frac{x-x_1}{a_2} + \frac{x-x_2}{a_3} + \dots,$$

where $a_k = v_k(x_k)$, and where, when the fraction is terminated after n divisions, the constant a_n is to be replaced by $a_n + (x-x_n)/v_{n+1}(x)$ in the last denominator. If we then set $x = x_k$, $0 \leq k \leq n$, the fraction terminates before the residual $(x-x_n)/v_{n+1}(x)$ is introduced. Since (4.3.7) is an identity, the result of replacing $1/v_{n+1}(x)$ by zero - that is, terminating the fraction with a_n - will give a function $R_{n+1}(x)$ which agrees with $f(x)$ at

the $n+1$ points, x_0, x_1, \dots, x_n , under the assumption that the constants a_0, a_1, \dots, a_n actually exist and that the portion of the truncated fraction preceding $x-x_k$ does not vanish when $x = x_k$, $k = 0, 1, \dots, n-1$.

If we introduce the notation

$$(4.3.8) \quad v_k(x) = z_k[x_0, x_1, \dots, x_{k-1}, x] = a_k,$$

reference to (4.3.5) gives

$$z_0[x] = f(x), \quad z_1[x_0, x] = \frac{x-x_0}{z_0[x]-z_0[x_0]}$$

$$z_2(x_0, x_1, x) = \frac{x-x_1}{z_1[x_0, x]-z_1[x_0, x_1]}$$

and in general

$$(4.3.9) \quad z_k[x_0, x_1, \dots, x_{k-1}, x] = \frac{x - x_{k-1}}{z_{k-1}[x_0, \dots, x_{k-2}, x] - z_{k-1}[x_0, \dots, x_{k-2}, x_{k-1}]}.$$

Accordingly, we have

$$(4.3.10) \quad z_k[x_0, x_1, \dots, x_{k-1}, x_k] = \frac{x_k - x_{k-1}}{z_{k-1}[x_0, \dots, x_{k-2}, x_k] - z_{k-1}[x_0, \dots, x_{k-2}, x_{k-1}]}$$

Thus the $z_k[x_0, \dots, x_{k-1}, x_k]$ is the k^{th} inverse divided difference of $f(x)$ relative to x_{k-1} and x_k . We shall refer to the quantity defined by (4.3.10) as a k^{th} inverted difference of $f(x)$. The inverted difference (4.3.10) is symmetrical in its last two arguments x_{k-1} and x_k . It is not generally symmetrical in its other arguments. Hence it must be formed from the specific inverted differences $z_{k-1}[x_0, \dots, x_{k-2}, x_{k-1}]$ and $z_{k-1}[x_0, \dots, x_{k-2}, x_k]$ which possess their first $k-1$ arguments in common. Hence the following calculational arrangement is most convenient:

$$\begin{array}{ccccccc}
 x_0 & f(x_0) & & & & & \\
 x_1 & f(x_1) & z_1[x_0, x_1] & & & & \\
 x_2 & f(x_2) & z_1[x_0, x_2] & z_2[x_0, x_1, x_2] & & & \\
 x_3 & f(x_3) & z_1[x_0, x_3] & z_2[x_0, x_1, x_3] & z_3[x_0, x_1, x_2, x_3] & & \\
 \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
 \end{array}$$

where the diagonal elements are the desired constants a_0, a_1, a_2, \dots which appear in (4.3.7)

Section 4.4 Thiele's Continued-Fraction Expansions

Whereas the k^{th} inverted difference, $z_k[x_0, \dots, x_{k-2}, x_{k-1}, x_k]$, of a function $f(x)$ was

symmetric in only its last two arguments, it happens that the quantity

$$(4.4.1) \quad \rho_k[x_0, \dots, x_k] = \sum_{\ell=0}^{[k/2]} z_{k-2\ell}[x_0, \dots, x_{k-2\ell}]$$

is symmetrical in all of its $k+1$ arguments. Here the last term on the right is $z_0[x_0]$ if k is even, and is $z_1[x_0, x_1]$ if k is odd. This quantity is often known as the k^{th} reciprocal difference of $f(x)$.

In particular, we have

$$\begin{aligned} \rho_0[x_0] &= z_0[x_0] = f(x_0) \\ \rho_1[x_0, x_1] &= z_1[x_0, x_1] = \frac{x_1 - x_0}{f(x_1) - f(x_0)} \end{aligned}$$

and calculation shows that

$$\begin{aligned} \rho_2[x_0, x_1, x_2] &= z_0[x_0] + z_2[x_0, x_1, x_2] \\ &= \frac{x_0 f_0 (f_1 - f_2) + x_1 f_1 (f_2 - f_0) + x_2 f_2 (f_0 - f_1)}{x_0 (f_1 - f_2) + x_1 (f_2 - f_0) + x_2 (f_0 - f_1)} \end{aligned}$$

in which case the symmetry is apparent.

Since (4.4.1) implies that

$$\begin{aligned} (4.4.2) \quad \rho_k[x_0, \dots, x_k] - \rho_{k-2}[x_0, \dots, x_{k-2}] \\ = z_k[x_0, \dots, x_k], \end{aligned}$$

reference to (4.3.9) shows that the successive reciprocal differences may be obtained by use of the recurrence relation

(4.4.3)

$$\begin{aligned} \rho_k[x_0, \dots, x_k] = & \frac{x_k - x_{k-1}}{\rho_{k-1}[x_0, \dots, x_{k-2}, x_k] - \rho_{k-1}[x_0, \dots, x_{k-2}, x_{k-1}]} \\ & + \rho_{k-2}[x_0, \dots, x_{k-2}] \quad . \end{aligned}$$

While this formula is less simply applied than (4.3.9), the symmetry of the k^{th} reciprocal differences permits its calculation from any two of the $(k-1)^{\text{th}}$ reciprocal differences having $k-1$ of its arguments in common, together with the $(k-2)^{\text{th}}$ reciprocal difference formed with those arguments. Hence a reciprocal-difference table may be constructed in the convenient form

$$\begin{array}{ccccccc} x_0 & f(x_0) & & & & & \\ & & \rho_1[x_0, x_1] & & & & \\ x_1 & f(x_1) & & \rho_2[x_0, x_1, x_2] & & & \\ & & \rho_1[x_1, x_2] & & \rho_3[x_0, x_1, x_2, x_3] & & \\ x_2 & f(x_2) & & \rho_2[x_1, x_2, x_3] & & & \\ & & \rho_1[x_2, x_3] & & & & \\ x_3 & f(x_3) & & & & & \end{array}$$

From this table we may determine the coefficients of (4.3.7) by combining (4.3.8) and (4.4.2) - a procedure

due to Thiele - so that

$$a_0 = f(x_0), a_1 = \rho_1[x_0, x_1], a_2 = \rho_2[x_0, x_1, x_2] - f(x_0)$$

$$a_3 = \rho_3[x_0, x_1, x_2, x_3] - \rho_1[x_0, x_1]$$

and so forth. Hence the required coefficients are formed from reciprocal differences appearing in the forward diagonal beginning with $f(x_0)$. Because of the symmetry, the data from the same table are available for the determination of formulae in which the ordinates are introduced in other arrangements, by choosing different paths made up of suitable contiguous diagonal segments. Each such expansion is identical with the one that would be obtained by the use of the inverted-difference array corresponding to an appropriate reordering of the abscissae, but only one array of reciprocal differences is needed for the formation of the entire set. Thus the use of reciprocal differences, rather than inverse differences, is generally advantageous only if several such formulae are required.

Section 4.5 The Error Term of Thiele Expansions

Let us examine the difference between

(4.5.1)

$$y(x_0, x_1, \dots, x_n, x) = y_0 + \frac{x-x_0}{a_1} + \frac{x-x_1}{a_2} + \dots \\ + \frac{x-x_{n-2}}{a_{n-1}} + \frac{x-x_{n-1}}{a_n}$$

which passes through the points (x, y) , $i = 0, 1, 2, \dots, n$, and

(4.5.2)

$$y(x_0, x_1, \dots, x_{n-1}, x, x) = y_0 + \frac{x-x_0}{a_1} + \frac{x-x_1}{a_2} + \dots + \frac{x-x_{n-2}}{a_{n-1}} + \frac{x-x_{n-1}}{a_n(x)},$$

where we use the notation $a_n(x)$ to remind us that (x_n, y_n) must be replaced by (x, y) .

Now, equation (4.5.2) is of no practical value for it shows merely that we can terminate the continued fraction to obtain an identity.

To obtain an error term we must find an expression for

$$\frac{p_{n+1}(x)}{q_{n+1}(x)} - \frac{p_n(x)}{q_n(x)}.$$

From difference equations (4.2.3) we have

(4.5.3)

$$\begin{aligned}
\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n} &= \frac{a_{n+1}p_n + b_{n+1}p_{n-1}}{a_{n+1}q_n + b_{n+1}q_{n-1}} - \frac{p_n}{q_n} \\
&= b_{n+1} \frac{p_{n-1}q_n - p_nq_{n-1}}{q_{n+1}q_n} \\
&= -b_{n+1} \frac{q_{n-1}}{q_{n+1}} \left(\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} \right) .
\end{aligned}$$

Now, since

$$\frac{p_0}{q_0} = \frac{a_0}{1} , \quad \frac{p_1}{q_1} = \frac{a_0a_1 + b_1}{a_1}$$

we have

$$\frac{p_1}{q_1} - \frac{p_0}{q_0} = \frac{b_1}{q_1} .$$

Using (4.5.3) we have in order

$$(4.5.4) \quad \frac{p_2}{q_2} - \frac{p_1}{q_1} = -\frac{1}{q_1q_2} b_1b_2$$

$$\frac{p_3}{q_3} - \frac{p_2}{q_2} = \frac{1}{q_2q_3} b_1b_2b_3$$

\cdot
 \cdot
 \cdot

$$\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n} = (-1)^n \frac{1}{q_nq_{n+1}} b_1b_2 \cdots b_{n+1}$$

If b_1 in these expressions is replaced by $x-x_{i-1}$ we see at once that

$$\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}$$

vanishes at the $n+1$ points x_i , $i = 0, 1, \dots, n$ - as is expected.

Let p_n/q_n in (4.5.4) be constructed from the set of points (x_i, y_i) , $i = 0, 1, \dots, n$ and let p_{n+1}/q_{n+1} be constructed from the set of points (x_i, y_i) , $i = 0, 1, \dots, n$ and $(X, f(X))$. Then

$$\begin{aligned} (4.5.5) \quad \frac{p_{n+1}(X)}{q_{n+1}(X)} - \frac{p_n(X)}{q_n(X)} &= f(X) - \frac{p_n(X)}{q_n(X)} \\ &= (-1)^n \frac{(X-x_0) \dots (X-x_n)}{q_n(X) q_{n+1}(X)}, \end{aligned}$$

where

$$q_{n+1}(X) = a_{n+1} q_n(X) + (X-x_n) q_{n-1}(X)$$

$$\text{and} \quad a_{n+1} = a_n(x_n, x_0, x_1, \dots, x_{n-1}, X)$$

$$= \frac{X - x_n}{a(x_1, x_2, \dots, x_{n-1}, x_n, X) - a(x_0, x_1, \dots, x_n)}$$

(4.5.6)

Since the calculation of q_{n+1} in (4.5.6) requires that we know $f(X)$, (4.5.5) cannot be employed in an error estimate. In practice, we would replace X by the nearest tabular point in calculating (4.5.6) and thus obtain an appropriate error estimate from (4.5.5).

Section 4.6 Thacher and Tukey's Algorithm for Obtaining Rational Interpolates

This algorithm gives, in the simple case, the same rational function as that given by the inverse divided-difference algorithm of section 4.2. However, Thacher and Tukey's algorithm does not require the generation of a large inverted divided-difference table to find the a_j of (4.3.7), and herein lies one advantage of this algorithm. A more significant point is that the algorithm can be generalized to obtain rational interpolants in which the degree of the numerator may exceed the degree of the denominator by k - where k is an integer - or vice versa. The generalization may be achieved in the following ways:

- (1) by using for the b_j in (4.2.1), functions that vanish at $x = x_{j-1}$
- (2) by choosing appropriate starting functions.

The algorithm may be derived as follows: The convergents, $p_i(x)/q_i(x)$, of the continued fraction

$$f(x) = a_0 + \frac{x-x_0}{a_1} + \frac{x-x_1}{a_2} + \dots$$

obey the relation (4.2.3) with b_i replaced by $(x-x_{i-1})$; that is

$$\begin{aligned}p_i(x) &= a_i p_{i-1}(x) + (x-x_{i-1})p_{i-2}(x) \\q_i(x) &= a_i q_{i-1}(x) + (x-x_{i-1})q_{i-2}(x) .\end{aligned}$$

Hence if we develop functions $t_i(x)$ that:

- (i) are zero for $x = x_i$,
- (ii) obey the above relations, and
- (iii) $t_{-1}(x)$ and $t_0(x)$ are given appropriate initial values,

we can find a_i , $i = 1, 2, \dots$ through use of

$$(4.6.1) \quad t_i(x_i) = a_i t_{i-1}(x_i) + (x_i - x_{i-1}) t_{i-2}(x_i).$$

If we give a_i the value

$$(4.6.2) \quad a_i = \frac{-(x_i - x_{i-1}) t_{i-2}(x_i)}{t_{i-1}(x_i)}$$

$t_i(x)$ may be found from

$$(4.6.3) \quad t_i(x) = a_i t_{i-1}(x) + (x - x_{i-1}) t_{i-2}(x)$$

and we can show that $t_i(x_i) \equiv 0$ as follows: Using (4.6.1) and (4.6.2) we have

$$\begin{aligned}t_i(x_i) &= \frac{-(x_i - x_{i-1}) t_{i-2}(x_i) t_{i-1}(x_i)}{t_{i-1}(x_i)} + (x_i - x_{i-1}) t_{i-2}(x_i) \\&= -(x_i - x_{i-1}) t_{i-2}(x_i) + (x_i - x_{i-1}) t_{i-2}(x_i) \\&\equiv 0 .\end{aligned}$$

To obtain the a_i , $i = 1, 2, \dots$, of the inverse divided-difference algorithm, we set

$$a_1 = \frac{x_1 - x_0}{f(x_1) - f(x_0)} = - \frac{(x_1 - x_0)(-1)}{f(x_1) - f(x_0)} .$$

Hence we have

$$t_{-1}(x) = -1 \quad \text{and} \quad t_0(x) = f(x) - f(x_0).$$

Employing (4.6.2) and (4.6.3) we can obtain all the remaining a_i by recursion.

For the generalization of the algorithm we shall follow Thacher and Tukey. The series of functions $t_i(x)$ may also be defined by

$$t_i(x) = q_i(x) f(x) - p_i(x) .$$

Provided that $f(x)$ is finite and that $q_j(x)$ does not vanish, a necessary and sufficient condition that

$$(4.6.4) \quad \bar{f}_j(x) \equiv p_j(x)/q_j(x) = f(x)$$

is that

$$(4.6.5) \quad t_j(x) \equiv q_j(x)f(x) - p_j(x) = 0.$$

Now suppose that for some integers $j > 0$ and $k \geq 0$,

$t_{j-2}(x)$ and $t_{j-1}(x)$ satisfy the conditions

$$(4.6.6) \quad t_{j-2}(x_i) = 0 \quad -k \leq i \leq j-2$$

$$(4.6.7) \quad t_{j-1}(x_i) = 0 \quad -k \leq i \leq j-1 \quad .$$

Then the functions

$$(4.6.8) \quad t_j(x) = a_j t_{j-1}(x) + (x - x_{j-1}) t_{j-2}(x)$$

and

$$(4.6.9) \quad t_j(x) = t_{j-1}(x) + \frac{x - x_{j-1}}{a_j} t_{j-2}(x)$$

vanish for $x = x_i$ ($-k \leq i \leq j-2$) since both $t_{j-2}(x)$ and $t_{j-1}(x)$ vanish at those points. (The simple case of the algorithm corresponds to $k = 0$ and the use of the above initial values of $t_{-1}(x)$ and $t_0(x)$ to obtain the same a_i as those found by use of the inverse divided-difference algorithm.) They vanish for $x = x_{j-1}$, since $t_{j-1}(x)$ and $(x - x_{j-1})$ vanish there, and they can be made to vanish at $x = x_j$, if we give a_j the value

$$(4.6.10) \quad a_j = -(x_j - x_{j-1}) t_{j-2}(x_j) / t_{j-1}(x_j) \quad .$$

A possibility of trouble comes when $q_j(x)$ has a root near one of the base points. Under this circumstance,

subtraction of nearly equal quantities can lead to appreciable losses of significant figures. We can avoid this difficulty by changing the recursion from (4.6.8) to (4.6.9) for one or more cycles. An additional possibility of difficulty occurs if $t_{j-1}(x_j) = 0$, leading to an infinite a_j . This can occur only if $p_{j-1}(x_j)/q_{j-1}(x_j) = f(x_j)$. Unless $f(x) \equiv p_{j-1}(x)/q_{j-1}(x)$ so that additional data are superfluous, postponing the introduction of x_j and $f(x_j)$ will resolve the difficulty.

Thus we have, subject to the conditions that no $q_j(x_i)$ vanishes for $i \leq j$, an iterative method by which, given a set of x_i and corresponding finite $f(x_i)$, and suitable starting functions, $t_{-1}(x)$ and $t_0(x)$, we may construct a sequence of functions $t_j(x)$ which vanish for successively larger sets of x_i .

Because of the linearity of (4.6.5), of the definition of $t_j(x)$, and of the alternative forms, (4.6.8) and (4.6.9) of the iterative algorithm, the functions $q_j(x)$ and $p_j(x)$ obey recursive relations of the same form, and with the same a_j , namely (4.2.3) with $b_n = x - x_{n-1}$, or if (4.6.9) is in use, rather than (4.6.8)

$$(4.6.11) \quad q_j(x) = q_{j-1}(x) + \frac{x - x_{j-1}}{a_j} q_{j-2}(x)$$

$$(4.6.12) \quad p_j(x) = p_{j-1}(x) + \frac{x-x_{j-1}}{a_j} p_{j-2}(x) .$$

Since on each application of the iteration the degree of $q_j(x)$ is no greater than the greater of (1) the degree of $q_{j-1}(x)$ and (2) one more than the degree of $q_{j-2}(x)$, and since a similar relation holds for the degree of $p_j(x)$, $p_{j-1}(x)$ and $p_{j-2}(x)$, the degree of either the numerator or the denominator of the rational interpolate will increase by at most one unit for every two iterations.

It is clear that the argument would apply equally well if any other function which vanishes for $a = a_{j-1}$ were substituted for the factor $(x-x_{j-1})$ in (4.6.8) or (4.6.9). However, except under rather special circumstances, such as where the function is known to be symmetric, so that the factor $(x^2-x_{j-1}^2)$ would be appropriate, such a modification does not appear to have particular merit for interpolation of functions of a single variable by ratios of polynomials.

Another possibility for increasing the flexibility of the algorithm lies in the choice of the starting functions. The only necessary restriction on the starting function is that $t_0(x_0)$ vanish. If $t_0(x_0)$ depends upon

more than one parameter, we may use the additional flexibility to make $t_{-1}(x)$ and $t_0(x)$ both vanish at k additional points, $x = x_{-1}, x = x_{-2}, \dots, x = x_{-k}$, where the value of k is determined by the particular starting functions selected.

Ordinarily, the full degree of possible generality is not necessary. However, the cases which follow are important:

$$(a) \quad q_{-1}(x) = q_0(x) = 1,$$

while $p_{-1}(x)$ is a polynomial of degree $k-1$, which agrees with $f(x)$ at $x = x_i$ ($-k \leq i \leq -1$) and $p_0(x)$ one of degree k , which agrees with $f(x)$ at $x = x_i$ ($-k \leq i \leq 0$) and

$$(b) \quad p_{-1}(x) = p_0(x) = 1,$$

while $q_{-1}(x)$ is a polynomial of degree $k-1$, which agrees with $1/f(x)$ at $x = x_i$ ($-k \leq i \leq -1$) and $q_0(x)$ one of degree k which agrees with $1/f(x)$ at $x = x_i$ ($-k \leq i \leq 0$), respectively. Case (a) ordinarily leads to a sequence of ratios of polynomials in which the degree of the numerator exceeds that of the denominator by k and $k-1$, alternately, while case (b) leads to a sequence in which the degree of the denominator exceeds that of the numerator by k and $k-1$, alternately.

All the rational functions so obtained will be minimal in that (neglecting the triviality of a numerical multiplier common to numerator and denominator) there are exactly as many free coefficients to be determined as there are points to be fitted. By a suitable choice of k and either case (a) or case (b), one can obtain a rational interpolant with any minimal combination of degrees for numerator and denominator. Since all such are unique (up to a common constant) all minimal rational interpolations can be obtained via the generalized algorithm.

The starting functions for cases (a) and (b) are clearly the interpolation polynomials for $f(x)$ and $1/f(x)$, respectively, based on the points x_{-k}, \dots, x_{-1} for $p_{-1}(x)$ or $q_{-1}(x)$, and on the point x_0 as well for $p_0(x)$ or $q_0(x)$.

If the explicit form of the interpolating function is desired, the starting functions must also be generated explicitly. The following algorithm allows this to be done in an iterative fashion.

Let $\psi_j^+(x)$ or $\psi_j^-(x)$ be the desired polynomial of degree $j+k$, $-k \leq j \leq 0$, depending upon whether we are

developing $p_j(x)$ or $q_j(x)$. Thus we must satisfy one of the following sets of conditions

$$(4.6.13) \quad \psi_j^+(x_i) = f(x_i) \quad \text{or}$$

$$(4.6.14) \quad \psi_j^-(x_i) = 1/f(x_i)$$

where $-k \leq i \leq j \leq 0$. Let $s_j(x)$ be a polynomial of degree $j + k + 1$ which vanishes for the x_i of (4.6.13) or (4.6.14). Then, if $\psi_{j-1}^\pm(x_i)$ satisfies (4.6.13) or (4.6.14), and $s_{j-1}(x_i)$ vanishes for $-k \leq i \leq j-1$, the functions

$$(4.6.15) \quad s_j(x) = (x - x_j) s_{j-1}(x)$$

and

$$(4.6.16) \quad \psi_j^\pm(x) = \psi_{j-1}^\pm(x) + \frac{[f(x_j)]^{\pm 1} - \psi_{j-1}^\pm(x_j)}{s_{j-1}(x_j)} s_{j-1}(x)$$

will satisfy the imposed conditions for $x = x_i$ ($-k \leq i \leq j$).

If we use as initial functions

$$(4.6.17) \quad s_k(x) = (x - x_{-k})$$

and

$$(4.6.18) \quad \psi_{-k}^\pm(x) = [f(x_{-k})]^{\pm 1}$$

we can generate our interpolation polynomial recursively. This algorithm, it may be noted, is subject to round-off error, and therefore is not as precise as some of the classical methods of generating interpolation polynomials explicitly. It has, however, the advantage of simplicity, and of adaptability to successive introduction of data points.

Once the starting functions, $q_{-1}(x)$, $p_{-1}(x)$, $t_{-1}(x)$, $q_0(x)$, $p_0(x)$ and $t_0(x)$ have been generated, the generalized algorithm proceeds exactly as in the simple case given below.

A streamlined version of the simple algorithm ($k = 0$) can be written as:

Starting Values

$$t_{-1}(x_1) = -1, \quad t_0(x_1) = f(x_1) - f(x_0)$$

$$\text{for } i = 1, 2, \dots, n$$

$$q_{-1}(x) = 0, \quad q_0(x) = 1$$

$$p_{-1}(x) = 1, \quad p_0(x) = f(x_0)$$

Iteration ($j = 1, 2, \dots, n$)

$$a_j = -(x_j - x_{j-1})t_{j-2}(x_j)/t_{j-1}(x_j)$$

$$\phi_j(x_i) = a_j \phi_{j-1}(x_i) + (x_i - x_{j-1}) \phi_{j-2}(x_i)$$

$$\text{or}^* \quad \phi_j(x_i) = \phi_{j-1}(x_i) + (x_i - x_{j-1}) \phi_{j-2}(x_i) / a_j$$

$$\text{for } i = j+1, j+2, \dots, n$$

$$\text{where } \phi_j(x_i) = t_j(x_i), q_j(x), \text{ or } p_j(x)$$

* Use one or the other for t_j , q_j , and p_j with a particular value of j .

Order of Calculation

$$a_j, t_j(x_i), q_i(x)^{**}, p_j(x)^{**}$$

** These two functions are calculated only if the rational function $p(x)/q(x)$ is required.

The calculation of the a_j - using the simple algorithm - is demonstrated by the following algorithm.

1. $i \leftarrow -1$
2. $t1(x_i) \leftarrow -1$
3. $t2(x_i) \leftarrow f(x_i) - f(x_0)$
4. $i \leftarrow i+1$
5. $i:n \xrightarrow{\leq} 2.$
6. $j \leftarrow -1$
7. $t2(x_j):0 \xrightarrow{=} 19$

8. $a_j \leftarrow (x_{j-1} - x_j) t_1(x_j) / t_2(x_j)$
 9. $i \leftarrow j+1$
 10. $i:n \xrightarrow{>} 16$
 11. $t_3 \leftarrow a_j t_2(x_i) + (x_i - x_{j-1}) t_1(x_i)$
 12. $t_1(x_i) \leftarrow t_2(x_i)$
 13. $t_2(x_i) \leftarrow t_3$
 14. $i \leftarrow i+1$
 15. Go to 10
 16. $j:n \xrightarrow{>} 21$
 17. $j \leftarrow j+1$
 18. Go to 7
 19. $m \leftarrow j-1$
 20. Go to 22
 21. $m \leftarrow n$
 22. Print: a_j for $j \leftarrow 1(1)m$
 23. Stop
 24. End
-

CHAPTER V

ALGORITHMS FOR OBTAINING RATIONAL APPROXIMATIONS

Section 5.1 Introduction

This chapter contains a few of the algorithms for obtaining rational approximations with emphasis on the ones that yield minimax - or nearly so - approximations.

The algorithms are presented in a descriptive manner, since the writer had to rely on incomplete references for the vast majority of the information contained in this chapter. It was not easy to uncover the assumptions used in these algorithms or to evaluate their efficiency on different types of computers. Thus consideration of these points has been omitted.

The problem of obtaining a minimax rational approximation, of a specified pair of degrees, to a continuous function is very difficult, because it is basically a non-linear problem. Some algorithms such as the extension of Remez's Second Algorithm and Maehly's Direct Methods attack the problem directly. But other algorithms, such as Maehly's Indirect Methods, telescoping procedure for continued fractions, and the Linear Inequality Algorithm of Loeb, attack the problem in a manner that leads to either a correction of the coefficients of a Padé or other rational approximant to give a lower order approximant that is nearly a minimax one, or transformation of the non-linear problem into a linear problem. Another problem - that of assuring an approximation that is free of poles in the interval of approximation - is overcome in many algorithms by restricting the denominator to values that are greater than zero for any value of x in the interval of interest.

The Padé method has been included, because Padé-approximants are often used as the initial approximants of Maehly's telescoping procedure for rational functions and Indirect Methods.

Section 5.2 The Padé Approximants

This method differs from the one given in Hall and Knight (see page 369 of [6]) in that it allows us to obtain rational approximations of varying pairs of degrees. On the other hand, the method given in Hall and Knight results in a rational approximation in a restrictive continued fraction form that provides convergents in which the degree of the numerator is alternately one greater than and equal to the denominator.

The Padé Table is a general method by which any power series can be transformed into a table of rational approximations to the function represented by the power series.

The coefficients p_k , q_k of a rational function $p_n(x)/q_m(x)$, where

$$p_n(x) = \sum_{k=0}^n p_k x^k$$

and

$$q_m(x) = 1 + \sum_{k=1}^m q_k x^k,$$

are deduced from the c_k of the power series $\sum_{k=0}^{\infty} c_k x^k$, for the function to be approximated, with the aid of the definition

(5.2.1)

$$q_m(x) \sum_{k=0}^{\infty} c_k x^k - p_n(x) = x^{n+m+1} \sum_{k=0}^{\infty} \gamma_k x^k.$$

This definition yields $n+m+1$ linear equations for $n+m+1$ unknowns $p_k, q_j, (k = 0, 1, \dots, n), (j = 1, 2, \dots, m)$, since $q_0 = 1$.

Padé approximations are most useful if $n = m$ or $n = m+1$.

Taking the general case we have from (5.2.1)

$$\sum_{l=0}^{\infty} c_l x^l \sum_{k=0}^m q_k x^k - \sum_{k=0}^n p_k x^k = \sum_{k=0}^{\infty} \gamma_k x^{m+n+k+1}.$$

Now,

$$\sum_{l=0}^{\infty} c_l x^l \sum_{k=0}^m q_k x^k = \sum_{l=0}^{\infty} \sum_{k=0}^{\min(l,m)} q_k c_{l-k} x^l$$

so that we get

(5.2.2)

$$\sum_{l=0}^{\infty} \sum_{k=0}^{\min(l,m)} q_k c_{l-k} x^l - \sum_{k=0}^n p_k x^k = \sum_{k=0}^{\infty} \gamma_k x^{m+n+k+1}.$$

Now, there is no term other than the first sum on the L.H.S. of (5.2.2) that has powers of x in the range $n+1 \leq l \leq n+m$. Hence, replacing l by $n+s$, $1 \leq s \leq m$,

we have

$$(5.2.3) \quad \sum_{k=0}^m q_k c_{n+s-k} = 0,$$

which we can write as a system of equations in the following way:

$$(5.2.3a)$$

$$\begin{bmatrix} c_n & c_{n-1} & \cdots & c_{n-m+1} \\ c_{n+1} & c_n & \cdots & c_{n-m+2} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ c_{n+m-1} & c_{n+m-2} & \cdots & c_n \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ \cdot \\ \cdot \\ \cdot \\ q_m \end{bmatrix} = \begin{bmatrix} c_{n+1} \\ c_{n+2} \\ \cdot \\ \cdot \\ \cdot \\ c_{n+m} \end{bmatrix}.$$

The lowest power of x on the R.H.S. of (5.2.2) is $n+m+1$.

Hence, we have

$$(5.2.4) \quad p_k = \sum_{\ell=0}^k q_\ell c_{k-\ell}$$

for $k = 0, 1, \dots, n$. The highest power of x in the second sum on the L.H.S. of (5.2.2) is n . Hence, letting $\ell = n+m+k+1$, we have

$$\gamma_k x^{n+m+k+1} = \sum_{i=0}^m q_i c_{n+m+k-i+1} x^{n+m+k+1}.$$

Hence,

$$(5.2.5) \quad \gamma_k = \sum_{i=0}^m q_i c_{n+m+k-i+1}.$$

Thus by solving the system of equations (5.2.3a) for the q_k , $k = 1, 2, \dots, m$, we can find the p_k , $k = 0, 1, \dots, n$, from (5.2.4).

The γ_k provide us with an error estimate. However, a crude idea of the error may be found from considering

$$|E_n| \leq |\gamma_0| x^{2n+1}.$$

For this, it is sufficient to compute γ_0 . γ_0 can be evaluated from the relation $\gamma_0 = \Delta_n / \delta_n$, where

$$\Delta_n = \begin{vmatrix} c_1 & c_2 & \cdot & \cdot & \cdot & c_{n+1} \\ c_2 & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ c_{n+1} & c_{n+2} & \cdot & \cdot & \cdot & c_{2n+1} \end{vmatrix}$$

and δ_n is the principal minor of Δ_n obtained by omitting the last row and column in Δ_n .

After obtaining the p_k and q_k , we can, if we wish, transform our rational approximation into an equivalent finite continued fraction.

The accuracy of this type of rational approximation decreases rapidly as the value of the argument increases because the expression for the error, $|E_n|$, usually depends much more on the smallness of the reduced range in which the approximations are used than on the factor γ_0 in the error term. Hence, unless we have a strongly convergent power series it is not advisable to use uncorrected Padé approximations. We will examine methods of correcting Padé approximants in section 5.4 and section 5.6.

Section 5.3.1 An Extension of Remez's Second Algorithm

This algorithm is an extension of Remez's algorithm for finding polynomial approximations to the determination of "best-fit" rational approximations.

It is helpful in describing the algorithm for minimax rational approximation to make reference to the simpler algorithm for polynomial approximation since many of the details are exactly the same. Defining $r(x) = f(x) - p_n(x)$, $a \leq x \leq b$, it is known (see Chapter II, section 6) that corresponding to the best polynomial approximation of degree n to $f(x)$, there exists a set of points x_i and a minimum deviation ρ such that

(5.3.1)

$$r(x_i) = (-1)^i \rho, \quad i = 0, 1, \dots, n+1$$

(5.3.2)

$$|r(x_i)| \geq |r(x)|, \quad a \leq x \leq b, \quad i = 0, 1, \dots, n+1$$

The determination of the set of points x_i , the minimum deviation ρ , and the polynomial $p_n(x)$ of best approximation is achieved by replacing the system of equations (5.3.1) and (5.3.2) by the following iterative scheme, in which the superscript j applies to the j^{th} stage of the iteration.

$$(5.3.3) \quad r^j(x_i^j) = (-1)^i \rho^j, \quad i = 0, 1, \dots, n+1$$

$$(5.3.4) \quad r^j(x_i^{j+1}) = \text{extremum of } r^j(x), \quad i = 0, 1, \dots, n+1$$

and for at least one i

$$r^j(x_i^{j+1}) = \max_{a \leq x \leq b} |r^j(x)|$$

The solution is started by choosing an arbitrary set of points $\{x_i^0\}$ and using the resulting linear system (5.3.3) to determine ρ^0 and $p_n(x)$. the maxima and minima (including the greatest extrema) of $r^0(x)$ become by (5.3.4) the set of points $\{x_i\}$ and a cycle of the computation is complete. In applications the process

can be terminated by a sufficiently close agreement between successive values of ρ or alternatively when there is sufficiently close agreement between the values of the $|r^j(x^{j+1})|$ and $|\rho^j|$.

As in the linear case, the minimization of

$$\rho(p_n, q_m) = \max_{a \leq x \leq b} \left| f(x) - \frac{p_n(x)}{q_m(x)} \right|$$

can be achieved by the minimization of an auxiliary function

$$\delta(p_n, q_m) = \max_{x \in X} \left| f(x) - \frac{p_n(x)}{q_m(x)} \right|$$

for an appropriate set X consisting of $n+m+2$ points.

The following algorithm systematizes the search for this set X . Given a trial set X , select p_n^0 and q_m^0 to minimize δ . If $\delta(p_n^0, q_m^0) \geq \rho(p_n^0, q_m^0)$ then (p_n^0, q_m^0) is a solution to the problem and X is the appropriate set. Otherwise, select a point ξ not included in X for which

$$\left| f(\xi) - \frac{p_n^0(\xi)}{q_m^0(\xi)} \right| = \Delta(p_n^0, q_m^0) .$$

Next, as described below, replace a certain element x of X by ξ and repeat the process with the new set X . It is always possible to do this in such a way that the values of δ increase.

We can modify the algorithm in the following way to give faster convergence. At the r^{th} stage we have a set of points $X_n = \{x_1^r, \dots, x_{n+m+2}^r\}$ that is used to generate a rational function $Q_{nm}^r(x)$. Then associate with each point x_j^r that extremum in its neighborhood nearest to x_j^r at which the error has the same sign as at x_j^r . Denoting this set of extrema by

$$T_r = \{\tau_1^r, \dots, \tau_{n+m+2}^r\}$$

we then let $X_{r+1} = T_r$, instead of replacing one x_j^r by that member of T_r at which the error has its greatest magnitude. The idea behind this is that, since the limiting set of points X corresponds to the extrema of $Q_{nm}^*(x)$ (the "best-fit" rational approximation), convergence of the algorithm should be more rapid using the above method of replacement.

Actually, we do not need a rational function $Q_{nm}^0(x)$ at all but only a set of points X_1 and a value of ρ which we would like to be as close as possible to X and ρ_{nm}^* , respectively. We can either use for X_1 the extrema of $T_{n+m+1}(x)$ - the Chebyshev polynomial of degree $n+m+1$, or the extrema of an economized approximation to $f(x)$ which generally also gives a good initial value for ρ .

The system of equations (5.3.1) is non-linear in the rational case. If we write them in the form

(5.3.5)

$$[f(x_1^1) - (-1)^1 \rho] \sum_{j=1}^m q_j (x_1^1)^j - \sum_{j=0}^n p_j (x_1^1)^j = -f(x_1^1) + (-1)^1 \rho,$$

$$i = 0, 1, \dots, n+m+1,$$

we see that the only non-linear factor is in the first term. Now, consider the system (5.3.5) split into two sets of equations, one of which we denote by S_{n+m+1} consisting of $n+m+1$ equations (which, for convenience is taken to be the first $n+m+1$ equations) and the other being the remaining last equation. The system S_{n+m+1} is linear in the unknowns $A = \{p_0, \dots, p_n, q_1, \dots, q_m\}$ and may be thought of as implicitly defining each member of A as a function of ρ . Therefore, the last equation may be thought of as a function of ρ alone and may be written as

(5.3.6)

$$\begin{aligned} F(\rho) = & [f(x_{n+m+1}^1) - (-1)^{m+n+1} \rho] \sum_{j=1}^m q_j (x_{n+m+1}^1)^j \\ & - \sum_{j=0}^n p_j (x_{n+m+1}^1)^j + f(x_{n+m+1}^1) - (-1)^{m+n+1} \rho = 0 \end{aligned}$$

The equation $F(\rho) = 0$ may be solved by the secant method, in which case the procedure is as follows:

(1) Choose two values of ρ , ρ_0 and ρ_1 and for each value solve the linear system S_{n+m+1} for $p_0, \dots, p_n, q_1, \dots, q_m$. Substituting these results into (5.3.6) we get $F(\rho_0)$ and $F(\rho_1)$.

(2) Using the secant method, get ρ_2 as

$$(5.3.7) \quad \rho_2 = \rho_1 + \frac{\rho_1 - \rho_0}{F(\rho_0) - F(\rho_1)} F(\rho_1)$$

(3) Solve S_{n+m+1} with $\rho = \rho_2$ and use (5.3.6) to get $F(\rho_2)$.

(4) Repeat steps (2) and (3) until the desired degree of convergence is achieved.

In practice it is convenient to apply (5.3.7) once only. This can be done because, when the algorithm is far from convergence, high accuracy in ρ is not worthwhile and, when the algorithm is near convergence, the values of ρ change very little from one stage to the next which enables one application of (5.3.7) to achieve high accuracy.

The key to the remainder of the computation is the search for T_r , the set of extrema of $f(x) - Q_{nm}^r(x)$. Ralston (see [20]) has used a search procedure with which he has searched for extrema of $f(x) - Q_{nm}^r(x)$ by starting at x_1^r and searching in the direction of increasing magnitude of error. When the peak is found, he then uses quadratic interpolation. By choosing a fairly coarse step size at the first stage and refining this as the algorithm converges, the total computation may be kept fairly near a minimum.

A proof of the convergence of this algorithm has been given by Ralston in [20] which appears to place non-restrictive conditions on $f(x)$, the function to be approximated. He proves the convergence of the algorithm by assuming the existence of a suitable initial approximation which provides the first extrema of $r^0(x)$ with the same sign as the first extrema of $r^*(x)$.

A program for this process in algorithmic notation is given in the appendix.

Section 5.3.2 The Loeb Weighted Minimax Algorithm

We wish to minimize the expression

$$(5.3.8) \quad \max_{a \leq x \leq b} |f(x) - Q(x)|,$$

where

$$Q(x) \equiv \frac{p_n(x)}{q_m(x)} \equiv \frac{\sum_{i=0}^n p_{n-i} x^i}{1 + \sum_{i=1}^m q_{m-i} x^i}$$

We may write (5.3.8) in the form

$$\max_{a \leq x \leq b} \frac{1}{q_m(x)} |f(x)q_m(x) - p_n(x)|$$

since $q_m(x)$ must not vanish in $[a, b]$ and hence $q_m(x)$ can always be made greater than zero throughout $[a, b]$.

We can now use the following iterative procedure: at the k^{th} step select $p_n^k(x)$ and $q_m^k(x)$ so as to minimize

$$\max_{a \leq x \leq b} \frac{1}{q_m^{k-1}(x)} |f(x)q_m^k(x) - p_n^k(x)|,$$

where the superscripts k indicate the stage of the iteration. In this subproblem, $1/[q_m^{k-1}(x)]$ is regarded as a weight-function. For practical purposes, the interval $[a, b]$ may be replaced by a discrete set $\{x_1, x_2, \dots, x_{n+m+2}\}$ - chosen in the same manner as the initial set of ordinates in section (5.3.1) - and the problem is thereby reduced to an over determined system of linear equations which is to be solved in the minimax or Chebyshev sense.

The advantage of this method is the linearity

of the system of equations which arise in each iteration step.

It has been reported by E. W. Cheney and T. H. Southard (see [4]) that the Space Technology Laboratories Inc. has used this algorithm with great success. For example, it is reported that five to ten steps of the iteration easily suffices to give best rational approximations with up to fifteen variable coefficients.

The conditions of $f(x)$ which are sufficient to guarantee the convergence of the above iterative process are not known.

Section 5.3.3 Loeb's Linear Inequality Method

This method, given by Loeb in [12], permits us to use linear programming techniques to obtain a solution to the problem stated in section 5.3.2, where the expression to be minimized was

$$\rho(p_0, \dots, p_n, q_0, \dots, q_m) = \max_{a \leq x \leq b} |f(x) - Q(x)|.$$

Now, if the number $\rho^* = \text{g.l.b.} \rho$ were known, the coefficients p_i, q_i could be obtained by solving the

the system of inequalities

$$| q_m(x)f(x) - p_n(x) | \leq \rho^* | q_m(x) |$$

for $a \leq x \leq b$. Hence, if we restrict ourselves to functions for which $q_m(x) \geq \alpha > 0$ in $[a, b]$, we may write the following system of linear inequalities

$$q_m(x) \geq \alpha$$

$$q_m(x)f(x) - p_n(x) \leq \rho^* q_m(x)$$

$$p_n(x) - q_m(x)f(x) \leq \rho^* q_m(x)$$

for $a \leq x \leq b$. The number ρ^* is not usually known at the beginning and has to be determined by successive trials. As an initial value of ρ^* , say ρ^0 , we could take the maximum deviation of a corrected Padé approximant to the function in the interval of interest. With the trial value, ρ^0 , the systems of inequalities may be tested for consistency. If it is inconsistent then $\rho^0 < \rho^*$; otherwise $\rho^0 \geq \rho^*$. It has been recommended that the consistency test be made by minimizing the auxiliary function

$$\delta(p_0, \dots, q_m) = \max_{a \leq x \leq b} \max[\alpha - q_m(x), |q_m(x)f(x) - p_n(x)| - \rho^0 q_m(x)]$$

The system of linear inequalities above is equivalent to the single inequality $\delta \leq 0$, and we may minimize δ by

linear programming.

Section 5.3.4 A Differential Correction Method

This method, which was given by Cheney and Loeb in [3], is effective without assumptions on $f(x)$ other than continuity. This does not imply that all of the other methods discussed require more restrictive assumptions. At the start of the process, $Q(x) \equiv p_n^0(x)/q_m^0(x)$ is arbitrary except for the condition $q_m^0(x) > 0$ in $[a,b]$. Now, assuming that $Q^k(x)$ has been determined, put

$$\rho^k = \max_{a \leq x \leq b} |f(x) - Q^k(x)|.$$

We then determine $Q^{k+1}(x)$ in such a way as to minimize the expression

$$\delta = \max_{a \leq x \leq b} [|f(x)q_m^{k+1}(x) - p_n^{k+1}(x)| - \rho^k q_m^{k+1}(x)]$$

under the restriction that the numbers

$$|p_0|, \dots, |p_n|, |q_0|, \dots, |q_m| \text{ be } \leq 1.$$

It has been shown (see [3]) that $\rho^k \downarrow \text{g.l.b.} \rho \equiv \rho^*$. The calculations to determine $Q^k(x)$ may be carried out by using the methods of "convex programming" since δ is a convex function of the coefficients, and the constraint set is also convex.

Section 5.4 Maehly's Telescoping Procedure for Rational Approximations

We will begin by examining a simple economization process of a polynomial, and then examine an economization procedure for rational approximants.

Let $f(x)$ be an analytic function which we wish to approximate on an interval $[x_1, x_2]$ to a given degree of accuracy. Let us assume that we have found an integer, n , such that the truncated power series,

$$(5.4.1) \quad p_{n+1}(x) = \sum_{k=0}^{n+1} c_k x^k,$$

is a sufficiently good approximation on our interval. We then wish to find a shorter polynomial $p_n^*(x)$ so that the condition

$$(5.4.2) \quad \max_{x_1 \leq x \leq x_2} |p_n^*(x) - f(x)| = E_p,$$

where E_p is the greatest lower bound of the error of the approximation, is at least approximately fulfilled. We may accomplish this by the "telescoping procedure" which can best be described if we write

$$(5.4.3) \quad x_1 = \epsilon u_1; \quad x_2 = \epsilon u_2$$

where $u_2 - u_1$ is of the order of one, while ϵ is a

parameter which characterizes the length of the interval $[x_1, x_2]$.

We next denote by $S_{n+1}(u)$ the polynomial which is uniquely defined by

$$(5.4.4) \quad S_{n+1}(u) = \sum_{k=0}^n S_k u^k + u^{n+1}$$

and

$$(5.4.5) \quad \max_{u_1 \leq u \leq u_2} |S_{n+1}(u)| = E_S,$$

where E_S is the greatest of the deviations of $S_{n+1}(u)$ from zero in $[u_1, u_2]$. $S_{n+1}(u)$ is essentially a Chebyshev polynomial.

The telescoping procedure consists of taking $p_n^*(x) = p_{n+1}(x) - c_{n+1}\epsilon^{n+1}S_{n+1}(u)$ where $u = x/\epsilon$. If the power series converges for $x_1 \leq x \leq x_2$, then

$$(5.4.6) \quad \begin{aligned} p_n^*(x) - f(x) &= [p_n^*(x) - p_{n+1}(x)] - [f(x) - p_{n+1}(x)] \\ &= -c_{n+1}\epsilon^{n+1}S_{n+1}(u) - \epsilon^{n+2} \sum_{k=0}^{\infty} \epsilon^k c_{n+2+k} u^{n+2+k} \end{aligned}$$

from which it follows that

$$(5.4.7) \quad \lim_{\epsilon \rightarrow 0} \frac{p_n^*(\epsilon u) - f(\epsilon u)}{\epsilon^{n+1}} = -c_{n+1}S_{n+1}(u).$$

The theory of Chebyshev polynomials tells us that $S_{n+1}(u)$ assumes the maximum of its absolute value, with alternating signs, exactly $n+2$ times within the interval $[u_1, u_2]$. Now (5.4.7) shows that this is also at least approximately true for $p_n^*(\epsilon u) = f(\epsilon u)$ if ϵ is sufficiently small; however this property characterizes the best-fit polynomial of degree n of our function $f(x)$ uniquely, according to a fundamental theorem of Chebyshev (see chapter II, section 6). Hence it is fair to say that (5.4.2) is at least approximately fulfilled if $c_{n+1} \neq 0$ and if ϵ is sufficiently small.

The telescoping procedure can be applied almost as easily to any type and form of continued fraction. Suppose that the function $f(x)$ which we wish to approximate can be represented by a convergent continued fraction of the form

$$(5.4.8) \quad f(x) = \frac{\alpha_0}{b_0} + \frac{\alpha_1 x}{b_1} + \frac{\alpha_2 x}{b_2} + \dots + \frac{\alpha_k x}{b_k} + \dots$$

and let us assume that the $(n+1)^{\text{th}}$ approximant, which can be written as

$$(5.4.9) \quad R_{n+1}(x) = \frac{\alpha_0}{b_0} + \frac{\alpha_1 x}{b_1} + \frac{\alpha_2 x}{b_2} + \dots + \frac{\alpha_n x}{b_n} + \frac{\alpha_{n+1} x}{b_{n+1}},$$

represents $f(x)$ well within the given error bounds on the interval $[x_1, x_2]$. Our wish is to find a shorter approximant, $R_n^*(x)$, so that the condition

$$(5.4.10) \quad \max_{x_1 \leq x \leq x_2} |R_n^*(x) - f(x)| = E_R,$$

where E_R is the greatest lower bound of the error of the approximation, is at least approximately fulfilled.

Let us first look at the uncorrected n^{th} approximant $R_n(x)$. From the theory of continued fractions it is known that

$$(5.4.11) \quad \lim_{\epsilon \rightarrow 0} \frac{R_n(\epsilon u) - f(\epsilon u)}{\epsilon^{n+1}} = c_{n+1} x^{n+1}$$

where c_{n+1} is given below by (5.4.16). Now (5.4.11) is fully analogous to (5.4.7). Hence we can hope to find corrections to the coefficients of $R_n(x)$ such that

$$(5.4.12) \quad \lim_{\epsilon \rightarrow 0} \frac{R_n^*(\epsilon u) - f(\epsilon u)}{\epsilon^{n+1}} = c_{n+1} S_{n+1}(u),$$

where u , ϵ , and $S_{n+1}(u)$ are as defined above.

Half of the constants in (5.4.8) and (5.4.9) are redundant; two continued fractions which have the same quotients r_k ,

$$(5.4.13) \quad r_k = \frac{\alpha_{k+1}}{b_k b_{k+1}}, \quad k \geq 0,$$

and the same α_0 , are exactly equivalent. Hence to obtain $R_n^*(x)$ we may alter either the α_k of the b_k of $R_n(x)$. In the first case we take

$$(5.4.14a) \quad R_n^{\alpha^*}(x) = \frac{\alpha_0^*}{|b_0|} + \frac{\alpha_1^* x}{|b_1|} + \dots + \frac{\alpha_n^* x^n}{|b_n|},$$

$$(5.4.14b) \quad \alpha_k^* = \alpha_k + \alpha_k',$$

$$(5.4.14c) \quad \alpha_k' = -\alpha_k S_k (-\epsilon)^{n+1-k} \prod_{m=k}^n r_m.$$

The corresponding formulae for correcting the b_k are:

$$(5.4.15a) \quad R_n^{b^*}(x) = \frac{\alpha_0}{|b_0^*|} + \frac{\alpha_1 x}{|b_1^*|} + \dots + \frac{\alpha_n x^n}{|b_n^*|},$$

$$(5.4.15b) \quad b_k^* = b_k + b_k'$$

$$(5.4.15c) \quad b_k' = b_k S_k (-\epsilon)^{n+1-k} \prod_{m=k}^n r_m$$

For finite values of ϵ the two forms yield slightly different numerical values, but both of them satisfy (5.4.12) in the limit $\epsilon \rightarrow 0$. It may also be shown that

$$(5.4.16) \quad c_{n+1} = (-1)^n \frac{\alpha_0}{b_0} \prod_{m=1}^n r_m.$$

An approximant of the form (5.4.8) can also be expressed as the quotient of two polynomials,

$$(5.4.17) \quad Q_n(x) = \frac{p_n(x)}{q_n(x)}$$

which can be computed from the recurrence formulae

$$(5.4.18) \quad \begin{aligned} p_m &= b_m p_{m-1} + \alpha_m p_{m-2} \\ q_m &= b_m q_{m-1} + \alpha_m q_{m-2} \end{aligned} \quad \text{for } m \geq 2$$

starting with

$$(5.4.19) \quad \begin{aligned} p_0 &= \alpha_0, & p_1 &= \alpha_0 b_1 \\ q_0 &= b_0, & q_1 &= b_0 b_1 + \alpha_1 x. \end{aligned}$$

The rational approximation satisfying (5.4.12) can be found from

$$(5.4.20a) \quad Q_n^{\gamma^*}(x) = \frac{p_n^{\gamma^*}}{q_n^{\gamma^*}} = \frac{p_n + \gamma_0 x + \sum_{k=2}^n \gamma_k p_{k-2}}{q_n + \gamma_1 x + \sum_{k=2}^n \gamma_k q_{k-2}}$$

with

$$(5.4.20b) \quad \gamma_k = -S_k(-\epsilon)^{n+1-k} \prod_{m=k}^{n+1} \frac{\alpha_m}{b_m}.$$

The last form, (5.4.20), is probably the most convenient for practical purposes, especially if the division time - of the computer - is more than twice the multiplication time.

Ralston, in [19], has given another method for the economization of rational approximations that, unlike Maehly's procedure, can be applied to all Padé approximations. The application - in some cases - of the process is simpler than that of Maehly's procedure, because Ralston uses the coefficients of the Chebyshev polynomial (of the first kind) to obtain the corrections to the coefficients of the approximation that give a minimax - or nearly so - rational approximation.

Section 5.5 Maehly's "Direct Methods" for Fitting Rational Approximations

The term "direct methods" indicates that the coefficients of the Chebyshev - approximant are computed directly. Maehly uses a weight function, $g(x)$, and thus his condition for the "best-fit" Chebyshev - approximant to a function, $f(x)$, is that

$$(5.5.1) \quad \max_{a \leq x \leq b} \frac{|Q_{nm}(x) - f(x)|}{g(x)}, \quad g(x) > 0,$$

be minimized, where $Q_{nm}(x)$ is the set of rational functions of the form

$$(5.5.2) \quad Q_{nm}(x) = \frac{p_n(x)}{q_m(x)} = \frac{p_n + p_{n-1}x + \dots + p_0x^n}{q_m + q_{m-1}x + \dots + q_0x^m}.$$

Each rational function $Q_{nm}(x)$ gives rise to an error curve, or error function, defined by

$$(5.5.3) \quad r(x) = \frac{Q_{nm}(x) - f(x)}{g(x)}$$

and we write the error curve of the "best-fit" approximant as

$$(5.5.4) \quad r^*(x) = \frac{Q_{nm}^*(x) - f(x)}{g(x)}.$$

An error curve is said to have standard form if it satisfies the following additional conditions:

(1) there are exactly $n+m+2$ critical points,

$$a = x_0^* < x_1^* < \dots < x_{n+m+1}^* = b.$$

(2) $|r(x)| = \rho$ at the critical points with alternating signs.

(3) $\frac{d}{dx} r^*(x)$ is continuous, and vanishes only for $x = x_i^*$, $i = 1, 2, \dots, n+m$.

A sufficient, but not necessary, condition that the error curve $r(x)$ be optimal, is that it has standard form. All the methods discussed in this section are based on the assumption that $r^*(x)$ has standard form.

The unique solution $r(x) = r^*(x)$, $x_i = x_i^*$ of the following set of equations

$$(5.5.5) \quad r(x_i) = (-1)^i \rho, \quad i = 0, \dots, n+m+1$$

$$\frac{d}{dx} r(x_i) = 0, \quad i = 1, \dots, n+m$$

or

$$(5.5.6) \quad Q_{nm}(x_i) - f(x_i) - (-1)^i \rho g(x_i) = 0, \\ i = 0, \dots, n+m+1$$

$$(5.5.7)$$

$$g(x_i) \frac{d}{dx} (Q_{nm}(x_i) - f(x_i)) - (Q_{nm}(x_i) - f(x_i)) \frac{d}{dx} g(x_i) = 0, \\ i = 1, \dots, n+m,$$

is supplied by every standard error curve and its critical points. There are $2(n+m+1)$ equations and the same number of unknowns. Hence the problem consists in solving the above system of nonlinear equations.

I. The First Direct Method

This is a two-stage iteration method which can be described as follows:

- (1) Make an initial guess at the critical points and solve the equation (5.5.6) to obtain a rational function $Q_{nm}(x)$ and a value of ρ .
- (2) A new guess at the critical points is then given by the extrema of $r(x)$, and this guess is used in the subsequent stage (1) step. It also solves the equations (5.5.7).

For stage (2) a simple searching procedure is recommended which is also used in the Second Direct Method and will be described later.

The determination of the polynomial equation for ρ poses problems, especially for large m ; also, the evaluation of the equation is, at times, difficult. It is also at times difficult to determine the value of ρ that provides a pole-free approximant. Hence use of the First Direct Method is not recommended.

II. The Transformation of Rational into Polynomial Approximation

In the polynomial case, stage (1) of the First

Direct Method is greatly simplified because the equation for ρ is linear. Hence, it is useful to note that fitting a rational approximant $Q_{nm}(x)$ can be reduced to fitting a sequence of polynomials of the same degree.

Choose two reference polynomials $\bar{p}_n(x)$ and $\bar{q}_m(x)$ such that the quotient $\bar{p}_n(x)/\bar{q}_m(x)$ is irreducible and at least one of them actually attains the degree n or m , respectively. The optimal error curve may then be written as

$$\begin{aligned}
 (5.5.8) \quad r^*(x) &= \frac{1}{g(x)} \left[\left(\frac{p_n^*(x)}{q_m^*(x)} - \frac{\bar{p}_n(x)}{\bar{q}_m(x)} \right) - \left(f(x) - \frac{\bar{p}_n(x)}{\bar{q}_m(x)} \right) \right] \\
 &= \frac{(p_n^*(x)\bar{q}_m(x) - q_m^*(x)\bar{p}_n(x)) - q_m^*(x)(f(x)\bar{q}_m(x) - \bar{p}_n(x))}{g(x)q_m^*(x)\bar{q}_m(x)}
 \end{aligned}$$

where $Q_{nm}^*(x) = p_n^*(x)/q_m^*(x)$ denotes the Chebyshev approximant of $f(x)$ with the weight function $g(x)$.

Clearly we may say that

$$(5.5.9) \quad r^*(x) = \frac{\psi_{n+m}^*(x) - F(x)}{H(x)}$$

is also the error curve of the polynomial

$$(5.5.10) \quad \psi_{n+m}^*(x) = p_n^*(x)\bar{q}_m(x) - q_m^*(x)\bar{p}_n(x)$$

with respect to the function

$$(5.5.11) \quad F(x) = q_m^*(x)(f(x)\bar{q}_m(x) - \bar{p}_n(x))$$

and the weight function

$$(5.5.12) \quad H(x) = g(x)q_m^*(x)\bar{q}_m(x).$$

$\psi_{n+m}^*(x)$ is the "best-fit" polynomial since $r^*(x)$ has standard form. Hence we may attack our problem by Remez's Second Algorithm, for example.

Since $q_m^*(x)$ is not known at the beginning of the iteration, we must start with a guess for $q_m^*(x)$. Then one step is taken in the direction of the "best-fit" polynomial $\psi_{n+m}^*(x)$. From the linear system

$$(5.5.13) \quad \psi_{n+m}^*(x) = p_n(x)\bar{q}_m(x) - q_m(x)\bar{p}_n(x)$$

we determine the new polynomials $p_n(x)$ and $q_m(x)$, which are approximations to $p_n^*(x)$ and $q_m^*(x)$, respectively. The new $q_m(x)$ enters the subsequent iteration step. The linear system (5.4.13) has $n+m+1$ equations and $n+m+2$ unknowns, but the quotient $p_n(x)/q_m(x)$ has in fact only $n+m+1$ free coefficients. Hence we may fix one of the coefficients of $p_n(x)$ or $q_n(x)$ and then the

system can be solved. It is important that the reference polynomial $\bar{q}_m(x)$ be positive in the interval $[a,b]$, and $\bar{p}_n(x)/\bar{q}_m(x)$ should not be a good approximation to $f(x)$; for, if it is, the formation of the difference $f(x)\bar{q}_m(x) - \bar{p}_n(x)$ may result in a substantial loss of accuracy.

III. The Second Direct Method

The error curve, $r^*(x)$, has exactly $n+m+1$ zeros, z_i^* , $i = 0, \dots, n+m$ in the interval $[a,b]$ if it is of standard form. Hence we may write

$$(5.5.14) \quad r^*(x) = G(x) \prod_{k=0}^{n+m} (x - z_k^*) .$$

If we characterize $r(x)$ by its zeros rather than by its extrema, we avoid the instability of the First Direct Method. This leads to the Second Direct Method which can be described as follows:

- (1) Letting $a < z_0 < \dots < z_{n+m} < b$ be a guess at the zeros of the optimal error curve $r^*(x)$, we obtain, through rational interpolation, an $Q_{nm}(x)$ which satisfies the conditions

$$(5.5.15) \quad Q_{nm}(z_k) = f(z_k) , \quad k = 0, \dots, n+m.$$

(2) The zeros of $r(x)$ are corrected by the calculated extrema. The corrected zeros \bar{z}_k , $k = 0, \dots, n+m$ are then used in the subsequent stage (1) step.

The main problem of this method is the correction of the zeros z_k using the extrema x_i . Maehly used the following method for correcting the zeros z_k which he obtained through use of a short variational argument (see page 265 of [16]). The correction, δz_k , to the zeros z_k can be found from the following system of equations

$$\sum_{k=0}^n \left(\frac{1}{x_i - z_k} - \frac{1}{x_0 - z_k} \right) \delta z_k = \ln \left| \frac{r(x_i)}{r(x_0)} \right|, \quad i = 1, \dots, n+m+1,$$

where we may replace

$$\ln \left| \frac{r(x_i)}{r(x_0)} \right| \quad \text{by} \quad 2 \frac{|r(x_i)| - |r(x_0)|}{|r(x_i)| + |r(x_0)|}.$$

Hence our system of equation becomes

$$\sum_{k=0}^n \frac{(x_0 - x_i) \delta z_k}{(x_i - z_k)(x_0 - z_k)} = 2 \frac{|r(x_i)| - |r(x_0)|}{|r(x_i)| + |r(x_0)|}, \quad i = 1, \dots, n+m+1$$

which is well-conditioned since it has its largest elements close to the diagonal.

In the determination of the extrema of the error curve $r(x)$ we cannot use Newton's method because the error curve inevitably carries "noise", which precludes the numerical computation of the derivatives $r'(x)$ and $r''(x)$. Hence a searching procedure was adopted which searched in equal distances for the largest or smallest value. As soon as a value $r(\bar{x})$ was found which surpassed its neighbors $r(\bar{x}-h)$, $r(\bar{x}+h)$ the searching was stopped and the three points were interpolated by a parabola with extremum \tilde{x} where

$$\tilde{x} = \bar{x} - \left(\frac{r(\bar{x}+h) - r(\bar{x}-h)}{r(\bar{x}+h) - 2r(\bar{x}) + r(\bar{x}-h)} \right) \frac{h}{2} .$$

The selection of the searching distance h must be given considerable care since the bulk of the computation in the direct methods consists of evaluating the function $f(x)$.

Section 5.6 Maehly's "Indirect" and "Combined" Methods for Fitting Rational Approximations.

The term "indirect methods" indicates that the coefficients of a given approximant are corrected by the addition of suitable quantities to give near "best-fit" approximants. The indirect methods described in this

section require the representation of a function $f(x)$, by a finite continued fraction, to the full accuracy required:

$$(5.6.1) \quad f(x) = \frac{\alpha_0}{|b_0|} + \frac{\alpha_1 x}{|b_1|} + \dots + \frac{\alpha_N x^N}{|b_N|} .$$

The function is to be approximated by a rational function $R_v^*(x)$, with $v < N$, in the Chebyshev sense. This, like the telescoping procedure for section 5.4, can be regarded as a problem of economizing continued fractions.

The original Indirect Method dealt with the case $N = v + 1$ but we shall discuss the Combined Methods, which deal with the general case $N > v$. Let us first describe the Combined Methods for polynomial approximation which are simple and may serve as an introduction to the more complicated rational function case.

I. The Combined Methods for Polynomial Approximation

Assuming that $f(x)$ is given as a high degree polynomial,

$$(5.6.2) \quad f(x) = \sum_{k=0}^N c_k x^k ,$$

within the interval $[0, b]$, we look for the polynomial $p_v^*(x)$ which approximates the function $f(x)$ best in the Chebyshev sense.

The n^{th} Padé-polynomial $\bar{p}_v(x) = \sum_{k=0}^v c_k x^k$ is a good approximant for $f(x)$ if b is sufficiently small. $p_v^*(x)$ is characterized by the error curve

$$(5.6.3) \quad r^*(x) = \frac{p_v^*(x) - f(x)}{g(x)}$$

having standard form. If we write

$$\Delta p_v(x) = p_v^*(x) - \bar{p}_v(x)$$

we may write (5.6.3) in the form

$$r^*(x) = \frac{\Delta p_v(x) - (f(x) - \bar{p}_v(x))}{g(x)}$$

which shows that $\Delta p_v(x)$ is the Chebyshev - approximant of $f(x) - \bar{p}_v(x)$ with respect to the weight function $g(x)$. A crucial point is to find a method to evaluate the difference $F(x) = f(x) - \bar{p}_v(x)$ without a substantial loss of accuracy.

Obviously, in the polynomial case, we have

$$F(x) = \sum_{k=v+1}^N c_k x^k .$$

Hence any direct method can be employed to determine a "best-fit" polynomial $\Delta p_v(x)$ to $F(x)$, which is then added to the Padé-approximant $\bar{p}_v(x)$.

II. A Formula for $f(x) - Q_v(x)$

If $f(x)$ can be represented in the form

$$f(x) = \frac{\alpha_0}{|b_0|} + \frac{\alpha_1 x}{|b_1|} + \dots + \frac{\alpha_N x}{|b_N|} ,$$

the combined methods consist in splitting off the n^{th} Padé-approximant

$$\bar{Q}_v(x) = \frac{\alpha_0}{|b_0|} + \frac{\alpha_1 x}{|b_1|} + \dots + \frac{\alpha_v x}{|b_v|} .$$

The crucial point is, as above, to find a formula for $\bar{Q}_v(x) - f(x)$ which can be evaluated without a substantial loss of accuracy.

Referring to the formulae derived in section 4.2 we may express our continued fraction as convergents, p_1/q_1 , which are obtained from the recursion relations

$$p_i = \alpha_i p_{i-2} + b_i p_{i-1}$$

$$q_i = \alpha_i q_{i-2} + b_i q_{i-1}$$

starting with

$$p_{-2} = 1, p_{-1} = 0, q_{-2} = 0, q_{-1} = 1.$$

Maehly, in the following manner, derived an expression for the difference $p_N/q_N - p_v/q_v$, $N > v$. Putting

$$f_{v+1,N}(x) = \frac{\alpha_{v+1}x}{|b_{v+1}|} + \frac{\alpha_{v+2}x}{|b_{v+2}|} + \dots + \frac{\alpha_{v+N}x}{|b_{v+N}|}$$

we may write

$$(5.6.4) \quad \frac{p_N}{q_N} = \frac{\alpha_0}{|b_0|} + \frac{\alpha_1 x}{|b_1|} + \dots + \frac{\alpha_v x}{|b_v|} + \frac{f_{v+1,N}(x)}{1}.$$

An expression for the difference of two successive convergents can be written as (see section 4.5)

$$\begin{aligned} \frac{p_{i+1}}{q_{i+1}} - \frac{p_i}{q_i} &= \frac{p_{i+1}q_i - p_i q_{i+1}}{q_{i+1}q_i} \\ &= \frac{1}{q_{i+1}q_i} \prod_{k=0}^{i+1} (-\alpha_k x) = \frac{1}{q_i(q_{i-1} + \frac{b_{i+1}}{\alpha_{i+1}} q_i)} \prod_{k=0}^{i+1} (-\alpha_k x). \end{aligned}$$

Applying this formula to (5.6.4) we obtain

$$\frac{p_N}{q_N} - \frac{p_v}{q_v} = \frac{1}{q_v(q_{v-1} + \frac{q_v}{f_{v+1,N}(x)})} \prod_{k=0}^v (-\alpha_k x) .$$

Hence, if we let

$$S_n(x) = q_{v-1} \quad \text{and} \quad q_m(x) = q_v ,$$

we obtain

(5.6.5)

$$f(x) - \bar{Q}_v(x) = \frac{(-x)^{v+1}}{q_m(x)(S_n(x) + \frac{q_m(x)}{f_{v+1,N}(x)})} \prod_{k=0}^v \alpha_k ,$$

which is the desired expression.

III. The Combined Method for Rational Approximations

The combined method described here arises from the First Direct Method. Essentially the method amounts to choosing polynomials $\bar{p}_n(x)$ and $\bar{q}_m(x)$, whose quotient is the Padé-approximant $\bar{Q}_v(x)$, as reference polynomials, and proceeding as in section 5.5.II. We recall that the error curve,

$$r^*(x) = \frac{Q_v^*(x) - f(x)}{g(x)} ,$$

was conceived, in that case, as the error curve of the polynomial

$$(5.6.6) \quad \psi_v^*(x) = p_n^*(x)\bar{q}_m(x) - q_m^*(x)\bar{p}_n(x)$$

approximating

$$F(x) = q_m^*(x) (f(x)\bar{q}_m(x) - \bar{p}_n(x))$$

with

$$H(x) = g(x)q_m^*(x)\bar{q}_m(x)$$

as the weight function.

Contrary to the direct approach of section 5.5.II, we insist that $\bar{p}_n(x)/\bar{q}_m(x)$ be a good approximation to $f(x)$. This is possible because formula (5.6.5) enables us to avoid separate evaluation of $f(x)$ and $\bar{q}_v(x)$. If we substitute the correction polynomials

$$\Delta p_n(x) = p_n^*(x) - \bar{p}_n(x), \quad \Delta q_m(x) = q_m^*(x) - \bar{q}_m(x),$$

into (5.6.6) we get

$$(5.6.7) \quad \psi_v^*(x) = \Delta p_n(x)\bar{q}_m(x) - \Delta q_m(x)\bar{p}_n(x)$$

from which $\Delta p_n(x)$ and $\Delta q_m(x)$ can be computed directly.

Since the equation (5.6.7) allows one degree of freedom, we normalize $p_n^*(x)$ and $q_m^*(x)$ by requiring that $q_m^*(x)$ have the same constant term as the Padé-polynomial $\bar{q}_m(x)$, which leads to the additional relation

$$(5.6.8) \quad \Delta q_m(0) = 0.$$

The method is again a two stage algorithm and can be described as follows:

(1) A guess at the critical points is made for which we determine

$$Q_v(x_i) = p_n(x_i)/q_m(x_i)$$

such that

$$Q_v(x_i) = (f(x_i) - \bar{Q}_v(x_i)) + (-1)^i \rho g(x_i),$$

$$i = 0, \dots, n+m+1.$$

The iteration runs as follows: make an initial guess at $q_m(x)$ and compute $\psi_v(x)$ and ρ from

$$(5.6.9) \quad \psi_v(x_i) = F(x_i) + (-1)^i \rho H(x_i).$$

We then determine $\Delta p_n(x)$ and $\Delta q_m(x)$ such that

$$(5.6.10) \quad \psi_v(x) = \Delta p_n(x) \bar{q}_m(x) - \Delta q_m(x) \bar{p}_n(x).$$

Finally, we replace $q_m(x)$ by $q_m(x) + \Delta g_m(x)$ and return to the beginning of the iteration.

(2) The extrema of

$$r(x) = \frac{\psi_v(x) - F(x)}{H(x)}$$

are used as a new guess at the critical points and stage (1) is re-entered.

As an initial guess at the critical points we use the critical points,

$$x_i = (1 - \cos \frac{i\pi}{n+m+1}) \frac{b}{2}, \quad i = 0, \dots, n+m+1,$$

of the transformed Chebyshev polynomial, if the interval $[0, b]$ is not too large. The initial guess at $q_m^*(x)$ is taken to be $\bar{q}_m(x)$, the denominator of the Padé-approximant $Q_v(x)$.

Maehly suggests the following iteration pattern since experience has shown that the initial guess at $q_m^*(x)$ is much poorer than the initial guess at the critical points. Starting with two or three steps of the stage (1) iteration, keeping the initial x_1 , we obtain a $q_m(x)$ close enough to $q_m^*(x)$ so the next stage (1) step may be combined with one step of the

stage (2) iteration. Thus the combined iteration step consists of determining $\psi_v(x)$ by (5.6.9), using the resulting error curve for correction of the x_i in stage (2), and finally correcting $q_m(x)$.

The method used for the determination of $\psi_v(x)$ by (5.6.9) was the following modification of Newton's interpolation method. Consider two polynomials $C_{v+1}(x)$ and $D_{v+1}(x)$ which satisfy

$$(5.6.11) \quad \left. \begin{aligned} C_{v+1}(x_i) &= F(x_i) \\ D_{v+1}(x_i) &= (-1)^i H(x_i) \end{aligned} \right\} i = 0, \dots, n+m+1.$$

We then have

$$\psi_v(x) = C_{v+1}(x) + \rho D_{v+1}(x)$$

where ρ is uniquely determined because the highest powers of $C_{v+1}(x)$ and $D_{v+1}(x)$ must cancel.

The two linear systems (5.6.11) can be solved simultaneously for the coefficients of $C_{v+1}(x)$ and $D_{v+1}(x)$ since the systems differ only with respect to their right-hand sides. The polynomials are written in Newton form,

$$C_{v+1}(x) = c_0 + c_1(x-x_0) + \dots + c_{v+1} \prod_{i=0}^v (x-x_i)$$

$$D_{v+1}(x) = d_0 + d_1(x-x_0) + \dots + d_{v+1} \prod_{i=0}^v (x-x_i) ,$$

in order to improve the condition of the linear systems. From this we obtain a triangular linear system with two right-hand sides for the coefficients c_i and d_i . We also have $\rho = -c_{v+1}/d_{v+1}$. Our resulting polynomial $\psi_v(x)$ again has Newton form,

$$\psi_v(x) = e_0 + e_1(x-x_0) + \dots + e_v \prod_{i=0}^{v-1} (x-x_i) ,$$

which guarantees stable evaluation of the error curve. However, $\psi_v(x)$ must be converted into the customary polynomial form for the computation of the corrections $\Delta p_n(x)$, $\Delta q_m(x)$.

Let us now consider a linear system for the correction polynomials. The coefficients of the reference polynomials $p_n(x)$ and $q_m(x)$, determine the coefficients of the linear system (5.6.10). This can pose serious problems, in most cases, because the coefficients of the reference polynomials display different orders of magnitude. Maehly solved the problem by triangularizing the system (5.6.10) in such a way that it can be solved without a substantial loss of accuracy. For a detailed description of the method used see pages 272-273 of [16].

CHAPTER VI

CONCLUSION

In this thesis an attempt has been made to employ basic Chebyshev concepts in the general case of approximation by rational functions.

It was seen in chapter III that the process of fitting a rational approximation, say $p_n(x)/q_m(x)$, to a prescribed set of $n+m+2$ points gives rise to $m+1$ values of $\min p$. At most one of the values of $\min p$ will give us an approximation that is free of poles in the interval of interest. A criterion that tells us, when we have a set of four ordinates, whether or not we can obtain a simple rational function of the form

$$\frac{a_0x + a_1}{b_0x + b_1} \quad ,$$

corresponding to either value of $\min p$, that has no pole in the interval of approximation, was given - in the form of a theorem - in chapter III. From the writer's experience with higher order rational approximations, it appears that the criterion can be generalized.

In chapter IV it was seen that it is not always possible to obtain an acceptable rational function that passes through a prescribed set of $n+m+1$ points and is of a specified form; that is, the function may have a pole in the interval of interest, or the system of equations for obtaining the coefficients of the rational function may be inconsistent. The fact that the system of equations may be inconsistent can raise difficulties in some of the algorithms of chapter V.

Of the various algorithms for obtaining interpolatory rational functions, Thacher and Tukey's method seems preferable to the others. The two main reasons for this preference are the generality of the algorithm and the fact that the algorithm does not require a large amount of storage for intermediate results when it is programmed for a computer.

The algorithms for obtaining minimax rational approximations, which are described in chapter V, can be considered as belonging to two classes; that is, methods that take the direct approach, and those that take the indirect approach. It appears that use of the methods of the first class will - in most cases -

provide a more nearly minimax approximation than the one obtained through use of the methods of the second class. However, most algorithms using the direct approach may require many iterations to arrive at a good approximation, especially if the initial approximation is poor.

It appears that an investigation of the problem of which combination of degrees of $p_n(x)$ and $q_m(x)$ - where n and m are variable, but their sum has an upper limit - leads to the least absolute value of $\min p$ and an approximation that has no poles in the interval of approximation would be of theoretical interest and considerable practical value.

BIBLIOGRAPHY

- [1] Altman, M.; Approximation Methods in Functional Analysis; California Institute of Technology, (1960).
- [2] Chebyshev, P. L.; Sur les questions de minima qui rattachement a la representation approximative des fonctions; Oeuvres Vol. I, 273-378, Chelsea, New York, (1961).
- [3] Cheney, E. W., and Loeb, H. L.; Two New Algorithms for Rational Approximation; Numerische Mathematik, 3, 72-75, (1961).
- [4] _____, and Southard, T. H.; A Survey of Methods for Rational Approximation SIAM Review, (July 1963).
- [5] Golomb, M.; Lectures on Theory of Approximation; Argonne National Lab., (1962).
- [6] Hall, H. S., and Knight, S. R.; Higher Algebra (Fourth Edition); MacMillan and Co., Limited, London, (1913).
- [7] Hamming, R. W.; Numerical Methods for Scientists and Engineers; McGraw Hill Book Co., Inc., Toronto, Ont., (1962).
- [8] Jackson, D.; American Math. Soc. Colloquium; New York, 11, (1930).
- [9] Kopal, Z.; Numerical Analysis; Chapman and Hall Ltd., London, (1955).

- [10] Lanczos, C.; Applied Analysis, Sir Isaac Pitmann and Sons, London, (1957).
- [11] Loeb, H. L.; A Note on Rational Function Approximation; Convair Astronautics Applied Mathematics Series, No. 27 (Sept., 1959).
- [12] _____; Algorithms for Chebycheff Approximations Using the Ratio of Linear Forms; J. Soc. Indust. Appl. Math., 8, 458-465, (1960).
- [13] Macon, N., and Dupree, D. A.; Existence and Uniqueness of Interpolating Rational Functions; Am. Math. Mon., 69, 751-758, (1962).
- [14] Maehly, H. J.; Rational Approximations for Transcendental Functions; I.B.M. publication RC - 86, (Jan., 1959).
- [15] _____; Methods for Fitting Rational Approximations, Part I: Telescoping Procedures for Continued Fractions; Jour. Assoc. Comput. Mach., 7, 150-151, (1960).
- [16] _____; Methods for Fitting Rational Approximations, Parts II and III; Jour. Assoc. Comput. Mach., 10, 257-277, (July, 1963).
- [17] _____, and Witzgall, C.; Tchebychev Approximations on Small Intervals, I and II; Technical report under Contract Nonr. 669 (14)(x), (August, 1960).

- [18] Pearson, K.; Tracts for Computers (No. II, Part I); Cambridge University Press, (1920).
- [19] Ralston, A.; Economization of Rational Functions; Jour. Assoc. Comput. Mach., 10, 278-282, (July, 1963).
- [20] _____; Rational Chebyshev Approximation by Remez' Second Algorithm; Stevens Institute of Technology (unpublished paper).
- [21] Rice, J. R.; Criteria for the Existence and Equioscillation of Best Tchebycheff Approximations; Jour. of Research of the National Bureau of Standards, 64B, 91-93, (1960).
- [22] Thacher, H. C., Jr., and Tukey, J. W.; Rational Interpolation Made Easy by a Recursive Algorithm; Argonne Nat. Lab., (March, 1961).
- [23] Todd, J., (edit.); Survey of Numerical Analysis; McGraw-Hill Book Co., Inc., Toronto, (1962).
- [24] Werner, H.; Ein Satz über diskrete Tchebycheff-Approximation bei gebrochen linearen Funktionen; Num. Math., 4, 154-157, (1962).

APPENDIX

Remez's Second Algorithm adopted to Rational Approximation

$$0. \quad c_1 \leftarrow .4$$

$$1. \quad k \leftarrow 1$$

COMMENT: Steps 2 to 23 set up the system of equations for p_i and q_i .

$$2. \quad i \leftarrow 1$$

$$3. \quad y(i) \leftarrow f(x_i) - (-1)^{i-1} \rho(k)$$

$$4. \quad i \leftarrow i+1$$

$$5. \quad i : n+m+2, \xrightarrow{\leq} 3$$

$$6. \quad i \leftarrow 1$$

$$7. \quad j \leftarrow 2$$

$$8. \quad M(i, j-1) \leftarrow Y(i) [(x_i)^{j-1}]$$

$$9. \quad j \leftarrow j+1$$

$$10. \quad j : m+1, \xrightarrow{\leq} 8$$

$$11. \quad i \leftarrow i+1$$

$$12. \quad i : n+m+1, \xrightarrow{\leq} 7$$

$$13. \quad i \leftarrow 1$$

$$14. \quad j \leftarrow m+2$$

$$15. \quad M(i, j-1) \leftarrow - (x_i)^{j-m-2}$$

$$16. \quad j \leftarrow j+1$$

$$17. \quad j : n+m+2, \xrightarrow{\leq} 15$$


```

18.       $i \leftarrow i+1$ 
19.       $i : n+m+1, \xrightarrow{\leq} 14$ 
20.       $i \leftarrow 1$ 
21.       $M(i, n+m+2) \leftarrow -Y(i)$ 
22.       $i \leftarrow i+1$ 
23.       $i : n+m+1, \xrightarrow{\leq} 21$ 
24.      Solve system of linear equations to
          obtain  $p_0, \dots, p_n$  and  $q_1, \dots, q_m$ . ( $q_0 = 1$ ).
25.       $i \leftarrow 1$ 
26.       $R(k, i) \leftarrow p(i-1)$ 
27.       $i \leftarrow i+1$ 
28.       $i : n+1, \xrightarrow{\leq} 26$ 
29.       $s(k, 1) \leftarrow 1$ 
30.       $i \leftarrow 2$ 
31.       $i : m+1, \xrightarrow{>} 35$ 
32.       $s(k, i) \leftarrow q(i-1)$ 
33.       $i \leftarrow i+1$ 
34.      Go to 31
35.       $k \leftarrow k+1$ 
36.       $k : 3, (<, =, >) \rightarrow (2, 37, 57)$ 

```

COMMENT: The following section of the algorithm gives the equations for $F[p(k)]$; ie., it evaluates $\sum_{i=0}^n p_i x_{n+m+2}^i$, and $\sum_{i=1}^m q_i x_{n+m+2}^i$, and $f(x_{n+m+2})$ and the products.

37. $k \leftarrow 1$
38. $i \leftarrow 2$
39. $ts(k,1) \leftarrow 0$
40. $ts(k,1) \leftarrow ts(k,1) + s(k,i)[(x_{m+n+1})^{i-1}]$
41. $i \leftarrow i+1$
42. $i : m+1, \xrightarrow{\leq} 40$
43. $ts(k,2) \leftarrow (-1)^{n+m+2} [1 + ts(k,1)]$
44. $i \leftarrow 1$
45. $ts(k,3) \leftarrow 0$
46. $ts(k,3) \leftarrow ts(k,3) + R(1,i)[(x_{n+m+2})^{i-1}]$
47. $i \leftarrow i+1$
48. $i : n+1, \xrightarrow{\leq} 46$
49. $ts(k,4) \leftarrow f(x_{n+m+1})ts(k,2) - ts(k,3)$

COMMENT: The equations for $F[\rho(k)]$ are given by
 $\rho(k)ts(k,2) + ts(k,4)$. The following finds
 $\rho(3)$ by the secant method.

50. $k \leftarrow k+1$
51. $k : 2, \xrightarrow{\leq} 38$
52. $E_1 \leftarrow \rho(1)ts(1,2) + ts(1,4)$
53. $E_2 \leftarrow \rho(2)ts(2,2) + ts(2,4)$
54. $\rho(3) \leftarrow \rho(2) + [\rho(2) - \rho(1)]E_2/(E_2 - E_1)$
55. $k \leftarrow 3$
56. Go to 2

COMMENT: The following section searches for T_r using the p_1, q_1 associated with $\rho(3)$.

$$57. \quad c_1 \leftarrow c_1/2$$

$$58. \quad c_1 : .015, \xrightarrow{\geq} 60$$

COMMENT: This gives c_1 a lower bound of .015.

$$59. \quad c_1 \leftarrow .015$$

COMMENT: Steps 60 to 75 give us a new x_1 if one exists.

$$60. \quad h \leftarrow 2c_1(x_2 - x_1)$$

$$61. \quad v \leftarrow \rho(3)$$

$$62. \quad xz \leftarrow x_1$$

$$63. \quad xz \leftarrow xz + h$$

$$64. \quad w \leftarrow f(xz) - Q(xz)$$

$$65. \quad w : v, \xrightarrow{\leq} 76$$

$$66. \quad xz \leftarrow xz + h$$

$$67. \quad vw \leftarrow f(xz) - Q(xz)$$

$$68. \quad vw : v, \xrightarrow{\leq} 72$$

$$69. \quad v \leftarrow w$$

$$70. \quad w \leftarrow vw$$

$$71. \quad \text{Go to 66}$$

$$72. \quad v : \rho(3), \xrightarrow{\leq} 75$$

73. Use inverse quadratic interpolation to find new x_1 .

$$74. \quad \text{Go to 76}$$

75. Use inverse linear interpolation to find new x_1 .

COMMENT: Steps 76 to 97 give a new x_{i-1} if one exists.

76. $i \leftarrow 3$
77. $h \leftarrow c_1(x_{i+1} - x_{i-1})$
78. $v \leftarrow \rho(3)(-1)^i$
79. $xz \leftarrow x_{i-1} + h$
80. $w \leftarrow f(xz) - Q(xz)$
81. $|w| : |v|, (<, =, >) \rightarrow (82, 85, 89)$
82. $xz \leftarrow x_{i-1} - h$
83. $w \leftarrow f(xz) - Q(xz)$
84. $|w| : |v|, (<, =, >) \rightarrow (86, 85, 88)$
85. $x_{i-1} \leftarrow (x_{i-1} + xz)/2$
86. $i \leftarrow i+1$
87. Go to 97.
88. $h \leftarrow -h$
89. $xz \leftarrow xz+h$
90. $vw \leftarrow f(xz) - Q(xz)$
91. $|vw| : |v|, \xrightarrow{\leq} 95$
92. $v \leftarrow w$
93. $w \leftarrow vw$
94. Go to 89
95. Use quadratic interpolation (with xz , $xz-h$, $xz-2h$ and vw , w , v respectively to find new x_{i-1}).
96. $i \leftarrow i+1$
97. $i : n+m+2, \xrightarrow{\leq} 77$

98. $h \leftarrow -2c_1(x_{n+m+2} - x_{n+m+1})$
99. $v \leftarrow \rho(3)(-1)^{n+m+1}$
100. $xz \leftarrow x_{n+m+2} + h$
101. $w \leftarrow f(xz) - Q(xz)$
102. $|w| : |v|, \xrightarrow{\leq} 113$
103. $xz \leftarrow xz + h$
104. $vw \leftarrow f(xz) - Q(xz)$
105. $|vw| : |v|, \xrightarrow{\leq} 109$
106. $v \leftarrow w$
107. $w \leftarrow vw$
108. Go to 103
109. $|v| : \rho(3), \xrightarrow{\leq} 112$
110. Use inverse quadratic interpolation to
find new x_{n+m+2}
111. Go to 113
112. Use inverse linear interpolation to find
new x_{n+m+2}
113. $w \leftarrow f(x_1) - Q(x_1)$
114. $i \leftarrow 2$
115. $v \leftarrow f(x_i) - Q(x_i)$
116. $|v| : |w|, \xrightarrow{\leq} 118$
117. $w \leftarrow v$
118. $i \leftarrow i+1$
119. $i : n+m+2, \xrightarrow{\leq} 115$
120. $|v| : \rho(3) + c_2, \xrightarrow{>} 123$

121. Print v , $p(1)$, $q(1)$
122. STOP
123. $\rho(1) \leftarrow \rho(3)$
124. $(5/4) \rho(3) : |v|, \xrightarrow{\leq} 127$
125. $\rho(2) \leftarrow |v|$
126. Go to 1
127. $\rho(2) \leftarrow (5/4) \rho(3)$
128. Go to 1
129. End

B29816